ISBN: 978-90-39357163

Printed by Uitgeverij Zuidam
Cover design by Marleen Birkhoff

“I think I know what you did last summer”

Improving data quality in panel surveys

“Ik denk dat ik weet wat je vorige zomer hebt gedaan”
Het verbeteren van datakwaliteit in panel surveys
(met een samenvatting in het Nederlands)

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof. dr. G.J. van der Zwaan, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op vrijdag 24 februari 2012 des ochtends te 10.30 uur

door

Peter Jaap Lugtig
geboren op 15 februari 1983
te Noord-Scharwoude

Promotoren:

Prof. dr. J. J. Hox

Prof. dr. G. J.L.M. Lensvelt-Mulders

Table of Contents

Acknowledgements

The thing I like most about doing research is doing research with others. Survey methodologists form a separate academic community, including their own conferences and journals. A lot of the theoretical frameworks that are used to explain what makes a good survey stem from psychology, socio-linguistics, and communication sciences. The analysis of survey data further involves a lot of statistics. Survey methodology therefore is a field of academic generalists, a field where I feel at home.

Many of the materials in this dissertation stem from discussions with other people. Since my first course in survey methodology, in September 2004, I have learnt a great deal from many people. I want to thank some people in particular.

I took my first steps in survey methods together with Daniel Oberski, Linda Bos and Meike Morren, who all went to pursue a Ph.D involving survey methods. The three courses by Willem Saris that I took at the University of Amsterdam were my inspiration to learn more about survey methods and do a Ph.D.

In 2006 I became employed at the department of Methods and Statistics at Utrecht University. Remco Feskens, Gerty Lensvelt-Mulders and Joop Hox have been supportive of my teaching and research ever since the beginning of my project. Apart from doing research on panel survey methods, I also taught many students. I want to thank them for sharpening my understanding of methods and statistics, and for a lot of inspiration. I also want to thank my colleagues; for the pleasant work environment, support and all our lunch conversations. In particular, I want to thank Kirsten Namesnik, Nijs Lagerweij, Irene Klugkist, Thomas Klausch, Rens van de Schoot, Chantal Molnar, Flip Boerland, Floryt van Wesel, Joris Mulder, Marieke Westeneng, Tina Glasner, Edith de Leeuw, Hans Landsheer, Ellen Hamaker, Jesper Tijmstra, Elly Korendijk and Ben Baarda.

For the chapter on mixed-mode surveys in this dissertation, I worked with Remco Frerichs and Assyn Greven from TeamVier B.V. in Amstelveen. I want to thank them for their thoughts and ideas on how to conduct mixed-mode surveys in practice, and for trusting me with their data.

The fourth chapter involved a lot of lonely hours with my favorite computer programmes; AMOS and MPLUS. Gerty Lensvelt-Mulders has been helpful all along, but I especially want to thank her for her support during my struggles in specifying the statistical models that were

necessary to decompose the effects on Dependent Interviewing under different levels of measurement error.

Chapter five originated at a meeting I had with Annette Jäckle at the European Survey Research Association's conference in Prague, 2007. Annette invited me to come and visit the Institute for Social and Economic Research at the University of Essex, which I finally did in 2009. I want to thank ISER, European Centre for Analysis in the Social Sciences (ECASS) for giving me the opportunity to stay in Colchester and learn a great deal from the survey methods people there. In particular, I want to thank Noah Uhrig, Annette Jäckle and Peter Lynn. I hope we will collaborate more in future. Earlier drafts of chapter 7 received helpful comments from Peter Lynn, Stephen P. Jenkins and Gerty Lensvelt-Mulders.

Writing about mixed-methods research to understand change over time, the topic of chapter six, proved to be enlightening for me. Mixing qualitative and quantitative studies was new for me, but the process of thinking and writing about mixed-methods made me realize there is a lot to gain when researchers leave the know track. I hope that in future, I will be able to employ more qualitative research methods to understand how respondent react to requests to participate in a survey and answer (repeated) survey questions.. I want to thank Rik Beerthuizen for conducting the qualitative interviews used in this paper. I also want to thank Joop Hox, Rens van de Schoot and Floryt van Wesel for comments on an earlier version of this paper.

For chapter 7, I would like to thank Joop Hox and Edith de Leeuw, Thomas Klausch and Anja Boevé and participants to the panel survey methods- and nonresponse workshops for their comments on earlier drafts. I would also like to thank Annette Scherpenzeel and CentERdata for providing me with data, and being supportive of my work throughout this project.

Finally, I want to thank all the people that surround me outside of work. My great friends, who remind me there is life outside of academia. And most importantly, Marleen, Wouter and my mother for always being there for me. My father passed away while working on this dissertation. I hope he would enjoy seeing the result.

Utrecht, 15 July 2011
Peter Lugtig

1 Introduction

The day Bob – the hero in this story - gets unemployed, his life changes. Not only will have less to spend each month, his shopping behavior may change. Daily routines change. Friendships or acquaintances with colleagues change. From a social scientist's perspective studying Bob from the moment he becomes unemployed is a very interesting endeavor. The changing behaviors and attitudes that Bob will undergo because of the fact he got fired from work can teach us a lot about how Bob, and people like him react to the life-changing event of becoming unemployed. Politicians and civil servants will also find Bob interesting. Will Bob find work again? How will he do so? And how does he cope with being unemployed. Will he become depressed, or can he cope well?

Following Bob, and similar people who undergo a major life-changing event is popular. Getting a detailed picture of what Bob experiences during his spell of unemployment can teach us a lot about the consequences of becoming unemployed for Bob. When more unemployed people are followed over time, we can also learn about unemployment in general. How problematic is the loss of income? Do people adapt themselves easily to their new life? What type of people will find a new job fast, and who won't?

It is these questions that social scientists and policy makes try to answers by using panel surveys. In panel surveys, the same people are followed over time to study change in individuals as well as the population. The most interesting changes in the life's of people like Bob occur quite shortly after the event of interest. Because it is often very difficult to set up a study design that is able to recruit Bob into the study right at the moment he hears the bad news from his employer, panel surveys often just follow a random selection of workers, and from that sample, study the people who get unemployed. Two other advantages of following people in general, is that we can collect data on Bob's life *prior* to his period of unemployment. Perhaps we find that Bob becomes depressed not after losing his job, but before losing his job. In that case, depression is not a consequence of becoming unemployed, it rather is the cause.

The study of causation is something like a golden grail to many scientists, and especially when experimental manipulation is not feasible, a panel survey is a great alternative as a research instrument. Panel surveys are becoming ever more popular. This may be, because of advancements in science and policy-making. Social science and policy studies therefore need to take into account more and more complex social developments. Developments in statistical software have enabled

empirical tests of these complex theories with data. To this end, there is a greater need for ever more, and more complex data that can be generated with panel surveys.

There are also several practical reasons why panel surveys are becoming more popular among substantive scientists. The first one is that more and more panel surveys are established, with few panel surveys being terminated. One of the first modern panel surveys was the Panel Survey of Income Dynamics (PSID), run by the University of Michigan in 1968. Other panel surveys that are devoted to the study of income are the Canadian Survey of Labour and Income Dynamics (SLID) and the Survey of Income and Programme Participation (SIPP) in the United States. In these surveys, households are the primary unit of investigation, as income is always shared within households. From these surveys followed more encompassing Household Panel Studies, that did focus on other aspects than income alone. The German Socio Economic Panel (GSOEP) started in 1985, the British Household Panel Survey from 1991, the Swiss Household Panel in 1999, the Household, Income and Labour Dynamics in Australia (HILDA) in 2001, and the Dutch Longitudinal Internet Studies in the Social sciences (LISS) from 2007 are all examples of this. In the last 15 years, new panel surveys follow specific groups of respondents for a specific reason. The elderly are for example studied in the Health and Retirement Survey (HRS) and the Survey of Health, Aging and Retirement in Europe (SHARE). Family Dynamics are studied. Once established, the costs of keeping these panel surveys going are relatively low. At the same time, the costs of doing a cross-sectional survey have swelled, mainly because of the increased efforts necessary to attain good response rates (de Leeuw & de Heer 2002).

This short overview leaves out all the patient-based panel surveys in psychology and health care, student monitoring systems, as well panel studies run by market research firms. It also leaves out cohort studies; a specific type of panel studies, where a specific group of homogeneous respondents are followed over time. They often start with a group of newborns, as was done in 1946, 1958, 1970 and 2000 in the United Kingdom.

Despite the research infrastructure, knowledge of the methodology – how to conduct panel surveys - remains underdeveloped. Much of what panel survey managers do has been based on anecdotal evidence, or methodological research from cross-sectional surveys. Although cross-sectional and panel surveys have many similarities, there are specific methodological challenges to doing a panel survey. This dissertation focuses on four examples of such methodological challenges. These four challenges form by no means an exhaustive collection of all issues arise in

setting up and conducting a good panel survey. A recent handbook, the "Methodology of Longitudinal Surveys", edited by Peter Lynn (Lynn 2009) provides a much more extensive overview. The collection of 5 studies in this book focuses on some of the issues that more recently have been problematic for methodologists, or prove. The four issues that are discussed are:

1. Panel surveys are nowadays often conduced using multi (more than one) or mixed interviewing modes (face-to-face, Internet, Telephone or mail). Are differences that we find in survey data collected in different interviewing modes due to differences in the people that take part in these surveys, or are they due to differences caused by the mode of interviewing? In other words: would respondent Bob give the same answer if he would be interviewed on the Internet or by phone?
2. Data from consecutive interviews should be measuring the same topic (be equivalent), in order to measure change in a good way. Complex constructs, which are measured using more than one survey question, are often not equivalent across time, making it impossible to compare estimates across change and study change. How can measurement equivalence be better understood using open interviews? In other words: would Bob's score on a depression scale be comparable with his answer one year later. And if not, how can we understand this?
3. One method to improve the collection of the same questions over time, is to use data from earlier interviews in later interviews. This procedure, called Dependent Interviewing is believed to increase data quality and lead to better survey estimates. But is the case? And which specific method of Dependent Interviewing leads to the best results?
4. Once people participate in a panel study, panel survey managements try to keep them into the study. In studying unemployment, it would for example be really bad to lose those people that decide to move in order to find a new job elsewhere. How can dropout from a study – or attrition – be studied, and what are the effects of attrition on data quality?

The next chapter introduces the overarching theme of this collection of studies. How can panel surveys be conducted in such a way that the data provide the best picture of what is truly going on in Bob's life or the lives of all the unemployed. In other words, what is data accuracy and data quality, and how can we study it? At the end of this chapter it should be clear how different design decisions in panel survey affect different aspects of data accuracy. Each study in this dissertation discusses one or

more of the specific methodological concerns in longitudinal surveys. Therefore, the outline of the rest of this book follows after the next chapter.

2 Chapter 2: Panel Survey data quality

This dissertation discusses five methodological studies that each study data quality in panel studies. Whether a survey is of high or low quality, depends on the perspective of the user, and can either either focus on the survey process, fitness for purpose, or survey errors (Biemer & Lyberg 2003). The survey process is usually the focus of panel funders and panel managements. It focuses on the establishment of survey standards, quality checks and documentation of the survey process. The 'fitness for purpose' perspective focuses on the user of survey data. The process that goes on behind the screens is for the user not so important, but data dissemination is all the more important. Biemer & Lyberg (2003) describe seven quality criteria for the end user: 1) the comparability of survey results (across time and place), 2) the coherence, 3) relevance, 4) accuracy of the data, 5) timeliness, 6) accessibility and 7) interpretability. Data accuracy is usually the focus of methodologists, and will also be the focus of this dissertation. How to assure and improve the accuracy of survey data?

2.1 Accuracy

Data that are accurate imply two things. They measure what they intend to measure or put in methodological terms: they are valid. Second, the estimates are precise, meaning that the amount of random errors for every respondent is small, or again in methodological terms, the data are reliable (Groves et al. 2004)

Validity and reliability are threatened by several sources of survey error. Some of these sources only affect the reliability or validity of survey data, while others affect both. The causes of survey errors are the same for panel and cross-sectional surveys, but panel surveys add a specific longitudinal component to each cause.

Excellent descriptions of these sources of error are to be found in Groves (2005) and Biemer & Lyberg (2003). These errors do not have to be related to each other. For example, the undercoverage of people without a phone in a Computer Assisted Telephone Survey, just leads to coverage error. Sometimes, the sources of error are however related. Respondents who only participate after one or several reminders to participate in a survey, seem to report with more errors (measurements), than respondents who participate after the first survey request (Tourangeau, Groves & Redline, 2010). In this case, the reduction in nonresponse error due to increased efforts to contact all potential respondents comes at the price of increased measurement error. Survey

methodologists have long realized this, at least since the publication of the concept of total survey error in 1978 (Groves 1978).

Survey errors have mostly been studied in cross-sectional surveys, and although panel studies and cross-sectional surveys bear many similarities, the longitudinal nature of data collection means that very specific longitudinal measurement errors can threaten data quality. This dissertation aims to add to the limited, but growing knowledge of such errors.

2.2 Longitudinal measurement errors

We know relatively little about longitudinal survey errors. Partly, this may reflect the fact that panel studies are not as ubiquitous as cross-sectional surveys, leading survey methodologists to concentrate on the more general sources of error in all surveys. Two edited books have focused on collecting evidence for the existence and size of longitudinal measurement errors. Whereas the volume edited by Kasprzyk, Duncan, Kalton & Singh (1989) was mostly exploring the existence of such errors, the volume by Peter Lynn could draw on a lot of empirical research conducted in the period 1989 to 2009. Evidence of the extent and conditions under which longitudinal measurement errors occur remain scarce. Even fewer studies have studied how different sources of longitudinal measurement errors interact. As a simple example, take different forms of nonresponse error. Nonresponse error occurs at the start of any panel survey, but it can be limited when survey organization put a lot of effort into convincing potential respondents to participate. What remains unknown, is whether people that need convincing to participate in the first wave of a panel survey, are more likely to participate in the second and subsequent waves. As survey organizations have only limited resources (in time, staff and money), should they devote their attention to recruit as many people from the sample in the first wave, or should they try to limit nonresponse in later waves if their goal is to achieve the highest possible data quality?

Recent research (Kaminska 2010; Tourangeau, Groves & Redline 2010) suggests that reluctant respondents provide data with more measurement errors than 'eager' respondents, who do not need incentives or additional conversion effort to respond in a survey.

The field of survey methodology is still a long way off from understanding the trade-offs between all possible types of measurement errors. This book seeks to add knowledge to some of the potential trade-offs, show how trade-offs between different measurement errors can be

studied, and finally, how certain new data collection methods may reduce measurement errors and increase data quality. Before discussing ways to assess measurement errors and trade-offs between them in panel surveys, the next section first describes the different types of measurement errors that exist. This chapter finishes with an outline of the rest of this book.

Coverage errors

Most of the coverage errors in longitudinal surveys occur in the sample selection process, and are as such no different than in cross-sectional surveys. One aspect that is particularly important is the decision of the panel survey mode. Will the fieldwork be conducted over the phone, the Internet, mail or by face-to-face contact? Phone and Internet do not cover the general population, but can be a good option when studying special populations, like students or employees. Most panel studies have traditionally relied on face-to-face as the primary interviewing mode, although the large household panel studies do use the other modes for contacting respondents and leaving drop-off questionnaires. More recently, the Dutch LISS study has tried to establish an Internet-panel by offering Internet to those households without access (Scherpenzeel & Das 2011). Because a sampling frame of e-mail addresses for the general population does not exist, survey researchers in future will have to rely on such mixed-mode designs to recruit panel members. Because face-to-face panel surveys are increasingly costly, survey organizations are under pressure to switch existing panel survey to mail, phone and Internet modes. In all these designs, coverage errors are likely to occur.

Another source of coverage error in panel surveys is the fact that the population of interest changes over the course of the study. It is common to add refreshment samples to long-running panels (Lynn 2009), but these usually consist of well-defined sub-groups who either dropped out, or were added to the study later on. An East Germany was added to GSOEP after unification (Wagner, Frick & Schupp 2006), while specific samples for Wales and Northern-Ireland were added to the BHPS (Lynn 2009; Taylor, Brice, Buck & Prentice-Lane, 2009). In the case of migrants it is far more difficult to add them to the sample of the panel survey. Although the amount of error in every wave because of population entries might be small, it can over time lead to accumulated coverage error. The reverse case, that original sample members move out of the population is less problematic, as in every wave, it can be verified that the respondent is still part of the target population. Still, respondents who migrate out of the population, are more likely to be reported ‘lost’ during

fieldwork process, amounting to nonresponse error, rather than coverage error.

Sampling error

The process of sampling error is essentially the same for longitudinal and cross-sectional surveys, although sampling error can be difficult to estimate due to geographical clustering, and refreshment samples that are added over the course of the panel study (Lynn 2009).

Nonresponse error

Initial nonresponse is problematic in panel surveys. The nonresponse rate in the first wave of a panel survey is the maximum response rate in any subsequent wave, assuming that the initial sampling units are not approached n every wave. Although there are some suggestions that panel surveys with low response rates in the first wave, suffer from lower attrition rates later on (Watson & Wooden, 2009), it is very important to achieve a high response rate at the start of a panel study.

Initial nonresponse rates have been increasing throughout the world (de Leeuw & de Heer 2002) in both cross-sectional and panel surveys. Biases resulting from high levels of nonresponse can be selective, but are not necessarily so (Groves & Peytcheva 2008).

One potential cause for differences in initial nonresponse between cross-section and panel surveys, is the fact that respondent are asked for consent to participate for a long period. Respondents who are put off by such a request, are likely to be different from those who do participate (Watson & Wooden, 2009). One solution to this would be to not put too much emphasis on the longitudinal nature of the study in the advance letter or first contact. The request for continuing participation can in fact be postponed until the end of the first interview. A positive interview experience, and the establishment of some rapport between interviewer and respondent, should then lead to higher consent rates to participate in wave 2 of the panel study.

Unfortunately, nonresponse rates in panel surveys keep accumulating after the first wave; a process that is called attrition. In wave 2, respondents who either refused explicitly, or did not want to insult the interviewer during the first interviewer drop out. After wave 2, nonresponse rates in panel surveys, typically remain low (Watson & Wooden 2009). Even very modest rates of nonresponse rate at every wave, can however lead to considerable attrition rates over the long run. Bias resulting from attrition is one of the most known and well-studied

types of specific longitudinal measurement error. Although the process of attrition is in many ways similar to nonresponse in a cross-sectional survey, there is one important difference. All respondents who drop out in a panel survey did at least participate in one wave of the study.

Adjustment error

Adjustment errors stem from the process of correcting survey data on the basis of fieldwork outcomes. Weighting and imputation are common methods to adjust the survey data, but these methods may introduce bias of their own. Lynn (2009; 2011) notes that the process to adjust survey data to initial selection probabilities and attrition by weighting becomes more and more intricate with every wave of the panel study. The longer a panel study spans, the more complex the sampling and nonresponse processes become. This in turn may yield either very large or very small weights for individuals in the study. On top of this, researchers seldomly want to use the entire panel to study substantive questions dating back to the start of the panel survey. They may either use a subset of the total population, or a subset of years from the entire panel study.

When either the population of interest or the time of the study is different from the panel survey, the standard weights that are computed by the survey organizations are unusable. Researchers then have to compute entirely new weights for their study, depending on the population of interest and the time span of the study. Computing weights is a highly specialized aspect of survey statistics, and not every applied researcher would be knowledgeable enough to do this. Although beyond the scope of this book, data imputation may be used more often than nowadays to account and adjust for errors due to nonresponse and sample selection procedures.

Measurement error

Among all possible survey errors, the reduction of measurement error has been one of the main goals of survey methodologists. Survey methodologists believe that measurement errors are easily amenable. Changing the wording of a question, or altering the visual design of a survey affect measurement errors associated with that question.

The causes for measurement errors are generally thought to stem from the question and answer process. Tourangeau, Rips & Rasinski (2000) describe four cognitive stages between the stimulus (question) and response (answer). First, the respondent tries to *comprehend* the survey question and the task required to complete this question. Second, the

respondent retrieves information from memory, and then thirdly *judges* these pieces of information to form an internal answer to the question. Finally, the internal answer is *reported* using the answer options designed by the survey researcher. Each of these stages may introduce measurement error: a respondent might misinterpret a question, fail to retrieve or judge the accurate information from memory. Finally, a respondent's internal answer may not correspond to the available answer options available. The twin aim of reducing measurement errors is to limit both systematic survey errors - for example due to general misunderstanding of a question - and random errors - for example caused by a response scale that does not offer enough detailed answering options.

The nature of the questions that are asked in longitudinal surveys often complicates the response process. Past events, or details of life changes are difficult to remember at all, let alone remember correctly. The type of events that are covered can also be complicated, further complicating the response process. Most household surveys for example ask details about the sources, timing and amounts of income since the last interviewer. It comes as no surprise that survey methodologists worry about the extent of measurement errors in such questions and in longitudinal surveys in general. Particularly, survey researchers have worried about measurement errors in retrospective questions, that lead to biased survey estimates and specific survey errors as the seam effect. The seam effect occurs when respondents are asked to date life events, or changes in their life. Seam effects occur when data from multiple waves in the panel study are combined, and show as a heap in the data around the time of the interview (the seam) (Callegaro 2008; Conrad, Rips & Fricker 2009).

When measurements in a longitudinal survey contain measurement error, this not only biases estimates based on the data from single waves. The problem of measurement error becomes larger when two consecutive measurements are compared to produce estimates of change. If "Bob" reports an annual income of €34000 at the first measurement, and €36000 at the second measurement, it would be tempting to conclude that his income has increased. When we know that the income questions in both waves contain measurement error, we cannot be certain whether this apparent change is a 'true' change. Perhaps the true score of "Paul" at both measurements was €35000? As the primary focus of panel surveys is to study change, it is very important to limit measurement errors in every single wave of the study.

More recently, survey methodologists have tried to use the fact that persons are repeatedly measured to specifically address the issue of measurement error in change estimates. One method of doing this is

using Dependent Interviewing (DI). DI is the process of providing respondent answers from a previous wave of the study to the current wave. In variations of Proactive Dependent Interviewing (PDI), respondents are presented their answer from the previous wave in the survey question. In the 'remind, continue' PDI-design, respondents are reminded of their earlier answer, but are then asked the normal (independent) survey question. In the 'remind, still' design, respondents are asked whether their old status is still the same, while in the 'remind, change' design the opposite question is asked: whether their status has changed (Jäckle 2009). As opposed to PDI, respondents in Reactive Dependent Interviewing (RDI) always have to answer the independent survey question first. Only when the data from the previous and current wave are inconsistent, or when a 'large' change occurs for quantitative variables do they receive feedback on their answer from the previous wave, after which they can adapt or explain their answer. The memory cue that PDI provides ideally alleviates the cognitive burden while respondents search for an answer, making it easier to complete the retrieval and judgment phase of the answering process (Tourangeau, Rips & Rasinski 2000).

Processing error

Processing errors occur when an answer by a respondent is incorrectly processed in the database containing all answers from all respondents. As a result, processing errors occur mainly in surveys where the answering-process is non-automated – when interviewers record answers or enter interview data in databases.

Computer Assisted Surveys, and self-interviewing greatly reduce processing errors. However, in some panel surveys, the use of interviewers is preferred over self-interviewing because of the complex nature of the survey. In such surveys, processing errors can and do occur. One method that simultaneously tries to limit measurement errors and processing errors is the use of so-called edit checks. With edit checks, data that are provided by the respondent can be both checked for consistency with other variables within the same interview, or the same data provided in earlier interviews (cross-wave). *Within-wave edit checks* use information collected earlier in the same interview to check the consistency of responses and detect potential reporting errors. Respondents are queried about sources they have not reported, but for which they are likely to be eligible, judging from responses given earlier in the interview (Pennell 1993). *Cross-wave edit checks* are specific to longitudinal surveys. They use information provided in previous interviews to check the longitudinal consistency of responses.

Respondents are queried about sources they have reported in the past, but not in the current interview (Jäckle 2009; Mathiowetz & McGonagle 2000) . In some surveys, such as the US Survey of Income and Program Participation, cross-wave edit checks are also used to verify whether apparent changes in amounts of receipt are genuine or due to a reporting or data entry error. Cross-wave edit checks are typically referred to as 'dependent interviewing' (DI) and we follow this convention. For simplicity, we refer to within-wave edit checks as 'edit checks'. Both Dependent Interviewing and edit checks simultaneously serve as controls to limit measurement errors and processing errors. They both can detect misreporting by the respondent, and misprocessing by an interviewer.

2.3 Trade-offs between survey errors

The different types of survey errors never occur in isolation. Survey errors are often thought to be linked, or to have 'a common cause' (Biemer 2010)

Studying links and trade-offs between survey errors is more difficult than studying the individual error sources separately. One specific 'common cause' of survey errors is caused by the mode of interviewing. Interviews may be conducted with an interviewer present or by self-interviewing and by telephone, computer, paper or in person. The use of showcards, touch-tone entry, further means that there are a multitude of ways in which interviews can be conducted. In the last decade, surveys have increasingly been conducted by mixing these modes simultaneously or consecutively. The mode of the survey however affects various survey errors. Coverage errors for example, will be different when respondents are contacted by telephone, or mail. Young people are less likely to have a landline phone than older people, while in a mail survey, almost everyone can be contacted. Nonresponse errors are likely to be larger in a mail survey than a telephone survey, while measurement errors in both modes are also likely to differ.

The mode-effect is a particular type of survey error that occurs in mixed-mode surveys (de Leeuw 2005). The mode-effect is mainly associated with differences in measurement errors between survey modes, but other differences in survey error naturally co-occur with the mode-effect. It therefore is difficult to establish whether there really is a difference in measurement errors between survey modes (mode effect). Should differences in substantive variables be attributed to the mode effect, or differences in the composition of the sample in both modes, caused by differences in coverage and nonresponse errors?

The second chapter of this book focuses on this issue, and introduces propensity score matching as a method to separate nonresponse errors from measurement errors. This method allows the evaluation of the existence, and extent of any mode-effect. Mixed-mode surveys are often cross-sectional surveys, but panel surveys too mix modes, to cut the costs of re-interviewing the same people at different times, or interview multiple people from the same family at lower costs (for example using drop-off questionnaires).

Apart from the trade-offs between survey errors in mixed-mode surveys, many other trade-offs are imaginable. For example, do respondents who attrite in a survey provide better or worse data (with less or more measurement error) than those who continue to be respondents?

2.4 Studying survey errors

Studying measurement errors is a difficult matter. Because we can never observe them directly, we always need additional information in order to evaluate measurement errors indirectly. If "Bob" is interviewed in a survey and answers that his annual gross income €36000, we cannot know whether this particular value is correct or incorrect. There are a number of ways to evaluate survey data however. Knowing Bob's true income would of course be very helpful to evaluate his data. This does not happen very often in survey research, and if it does, there is usually no reason to also ask for these data. Before discussing a number of ways to use other information to study measurement errors indirectly, it is first necessary to discuss how measurement errors can affect substantive survey estimates.

Generally, measurement error can affect estimates in two ways. First, survey errors can lead to a systematic bias that will affect survey estimates. For example, if nonresponse bias is systematic, and lower educated respondents are underrepresented in the survey data, this may systematically bias all survey estimates related to education. Systematic survey errors may affect for example the distribution of variables, or substantive statistics, like the mean and median. For this reason, survey methodologists are concerned about estimating and correcting for these errors. It is even better however to make sure these systematic errors are reduced or eliminated by specific survey design features. Already noticed by Groves (2005), there seems to be a divide between those people who try to 'reduce' errors and those who try to 'measure' it.

Random errors cancel each other out within a sample. One such error is sampling error, but measurement and processing errors are to a large extent random as well. As such, random errors do not affect estimates like the mean and median, but only affect variances. Although survey methodologists try to reduce random survey errors, it is almost impossible to prevent random errors from occurring at all.

Validation data

The easiest and best way to study the presence of survey errors in survey data, is to compare them against true values of those data. Validation data at the respondent level are a very rich source for studying survey errors.

Validation data can sometimes be obtained through government records. In the case of health variables, survey measures can be validated against bio-physical measures. For the majority of variables, be they attitudes, behaviors and facts, validation data will never be available. For such variables, we need specific study designs, or statistical models to study measurement error.

Experiments

Survey experiments are a survey methodologist's best friend. When two versions of a survey question can be randomly administered to a sample of respondents, any difference between the two versions of the questions will be the result of the experimental manipulation (or sampling error). In both questionnaire design (Groves et al. 2004) and the visual design of (web) surveys (Dillman 2007; Smyth 2006; Toepoel 2008) experiments have taught a lot about what works and what does not. Not everything can as easily be manipulated as a question however. Some survey design features are not easily randomized, because they are costly (i.e. two or more sampling procedures) or because they are not easy to manipulate (nonresponse and processing error).

The second problem is that it often remains unclear what the effect of the experimental manipulation is on survey error. Finding a difference between two

versions of a survey question is not enough to evaluate the effect on data accuracy. Indirect evidence on the effect of survey experiments on survey errors can be found by linking the survey data to other data (testing the construct validity), and sometimes it is possible to define ex ante which survey design works best. For example, we know from practice that the use of soft and hard drugs remains underreported in surveys. If

experimental versions of questions are tested, we can assume that the version that generates the highest prevalence rate of drug use, is the best one.

A complicating factor is that the different survey errors interact. Experimenting with survey questions may have unintended consequences. In an experiment of different survey questions by Glasner (2011), she found that respondents dropped out of a survey more often in one of the conditions of the Event History Calender she implemented in a panel survey. In her experiment, it remained unclear whether this higher dropout rate in one condition also led to increased nonresponse bias, but it may have occurred. Similarly, a very detailed question about a respondent's profession, may lead to less measurement error than a more general question. However, the more detailed questions may also lead to more processing errors, and the net effect on survey errors may therefore be none.

Experiments in general are so powerful, because differences in the outcome variable can be attributed to the experimental manipulation. In the case of survey experiments, that is however not always the case, because the experimental manipulation often affects differences types of survey errors simultaneously. In order to separate the effect of experiments on the different types of survey error, as well as to study the net effect of all these survey errors, we need statistical models.

Statistical models to study survey errors

Rather than identifying the amount of survey error for every respondent, statistical models that have been developed to study survey errors, only estimate the aggregate amount of systematic error and random errors for a specific variable.

The goals of these statistical models is to separate one type of survey error from another one, thus making it possible to analyze the effects of experiments with survey characteristics, or assess the extent of survey error for groups of respondents. All these statistical models try to isolate specific sources of measurement errors, but they do so in different ways. Many of these models are extensions of a simple regression model. In experiments, the experimental manipulation can be included as a separate parameter in a regression model. In the simplest form of statistical modeling, the difference in the mean between two experimental conditions can show up as a dummy indicator in the regression model.

When the experimental manipulation is believed to affect not the mean, but rather the variances and covariances, the regression model is easily extended to a multi-group model. In this model, the experimental manipulation is the grouping variable, and differences caused by the experimental manipulation can show up in the constant, the regression coefficients or the mean square error (Groves 2005).

Multiple indicators

The models above rely on a regression model, and only include one observed measure for every variable in the model. Most of the more sophisticated models that have been developed to study measurement errors use multiple indicators for measurement the same variable (often called construct). Two approaches are possible: 1) measure the same construct twice in the same survey using slightly different questions (called parallel measures), or 2) the same individual survey questions can be measured more than once among the same people in different surveys. It should come as no surprise that the models that study measurement error in this book use the second approach. I therefore focus on the second approach. For the use of parallel measures, I refer to the excellent book of Alwin (2007).

Longitudinal models

Longitudinal models to study measurement errors all use repeated observations for the same people to estimate measurement errors.

One specific model that has been exclusively designed to study measurement errors, is the quasi-simplex model (Alwin 2007; Heise 1969; Wiley & Wiley 1970). The quasi-simplex model can be used to study the reliability of individual survey questions, when the question is measured three times or more among the same people. A more technical description of the quasi-simplex model follows in chapter 3. The basic idea of the model is that the correlations between all measures can be used to estimate what is the 'true' correlation between measures, and hence, also estimate the random measurement error at each observation. As the true score consists of both the observed score and measurement error, we can estimate the proportion 'true score' variance compared the observed score variance. This proportion equals the reliability coefficient (Alwin 2007).

A special case of the longitudinal model to study measurement errors is the Multi-Trait Multi-Method model (MTMM) as developed by Campbell

and Fiske (1959). The underlying idea of MTMM models is that they use both parallel measures and repeated measures to decompose variances. MTMM models have been extensively used to study how for example a question's response scale or mode of administration influence the validity and reliability of a survey question. The MTMM models has been used frequently over the past 50 years (see the special issue of Methodology (Eid 2009)) despite the practical and analytical difficulties that are associated with them. Alwin (2007) points to the problem that in MTMM models, the same question is repeated two or three times within a survey. Although many respondents will forget their answers in between, the risk of correlated answers or measurement errors seems high.

Statistical models to correct for survey errors

The models we discussed above do not necessarily attempt to correct for measurement errors. Their main goal is to estimate the presence of measurement errors in specific survey questions, with the goal of evaluating and improving survey questions. Longitudinal and parallel measures models can however be easily combined with substantive statistical models by using latent variables. In such models, the measurement error is separated from the true score of every variable, and then only the true scores are connected to the (true) scores other substantive variables. Such hybrid Structural Equation Models (SEM) do attempt to correct for biases due to measurement errors that may bias the covariances between variables. When surveys intend to compare groups, multiple-group analysis can be used to test the equivalence of the parameter estimates across the groups. This approach is feasible as long as researchers are interested in group differences and measurement errors.

When differences between groups are however caused by more than measurement errors alone, the multi-group SEM approach can not be used. For example, when measurement errors in two groups of respondents are caused by both differences in measurement error and differences in nonresponse error, the statistical models described above can only produce an estimate of the presence of the total difference between the groups. When researchers do not know whether group differences are caused by measurement errors or nonresponse errors, it remains difficult to improve survey procedures to limit or prevent such errors.

Weighting

The common method in surveys to deal with survey errors due to nonresponse is to weight the data (Bethlehem & Keller 1987). If men, for example, are more likely no to respond in a survey than women, the data of all mean are given a greater weight during data analysis than women. After weighting, the differences due to nonresponse between men and women have been corrected for. Weights are routinely computed for demographic variables and often included in the publicly available datasets that result from surveys. The use of weights remains limited to those variable for which information on both respondents and nonrespondents is available. This means that weighting can only partially separate the different sources of survey error.

Also, weighting may improve data quality and reduce the total amount of survey error, but it does not necessarily so. The effect of weighting on survey error ultimately depends on how well the weighting variables explain the mechanics that lead to nonresponse (Bethlehem & Keller 1987; Lee 2006).

Matching

The similarity between matching and weighting is that both methods can be used to correct for specific forms of survey errors. The main difference between matching and weighting is how the methods compare groups. With weighting, differences between two (or more) subgroups in the sample are corrected.

In matching, pairs of respondents from different groups (e.g. respondents and nonrespondents) are linked to each other based on the similarity of their background statistics (Deheji & Wahba 2002). In a simplified example, nonresponding males with different levels of education are matched to one or more counterparts from the respondents with similar backgrounds. People who cannot be matched can then be compared against people who are matched, while the matched respondents from both samples can also be compared to see whether after matching, differences on other variables remain.

The technique of matching is not frequently used by survey methodologists to study survey errors. Like weighting, an important assumption in matching is that the covariates that are used to match people, actually explain the differences between the two subgroups which are compared (e.g. nonrespondents and respondents). Also like weighting, the covariates should also be related to the variable of interest. In cross-sectional surveys, it is unlikely that one set of covariates

serves both these purposes well. One advantage of longitudinal surveys however, is that after the first wave of data collection, it is possible to use data from survey interviews, to correct for survey errors that are introduced later in the survey. Therefore, both weighting and matching can be used effectively in panel surveys.

2.5 Outline

In the following five chapters, several methodological innovations in panel surveys are evaluated. In each chapter, one of the methods discussed above to study and correct for measurement errors will be used to study how these methodological innovations affect survey errors and/or substantive conclusions derived from these survey data. The techniques discussed in the different chapters all build on one or more of the basic methods, but describe and explore the techniques in far more detail.

In Chapter 3, the technique of propensity score matching is used to study the effects a mixed-mode respondent recruitment strategy for a survey. It shows how matching can be used to separate nonresponse error from measurement error in a mixed telephone and Internet survey. Separating the two enables us to study how differences between the samples that remain after correcting for nonresponse error persist: the mode effect.

In Chapter 4, we turn to the technique of Dependent Interviewing (DI). Different versions of DI are experimentally compared and evaluated using a quasi-simplex model. This chapter shows how DI and the extent of measurement error present in a survey question on income affects the reliability coefficient.

Chapter 5 further explores the use of Dependent Interviewing in panel surveys. This chapter focuses on the effect DI has on substantive estimates that use income questions. Apart from this, details of a validation study using the same income questions shed light on how DI works to affect survey estimates.

Chapter 6 focuses on the topic of change in attitude question in a population that experiences a period of life changes. A mixed-method study that combines longitudinal survey data with qualitative interviews shows how attitudes change over time. Not only do levels of attitudes towards their study change among a group of first year psychology students, the concept of interest itself also changes. The chapter shows how the meaning of study motivation for students itself changes over time.

The final chapter focuses on panel attrition. Recent advances in mixture Structural Equation Modeling are used to describe the process of attrition in a panel study with monthly measurements. The chapter shows how different archetypes of respondents drop out of a study in different ways and for different reasons. This chapter concludes by showing how every group of attriters affects longitudinal nonresponse error in a different way.

3 Chapter 3: Estimating nonresponse bias and mode effects in a mixed-mode survey[1]

It is becoming more difficult and costly to conduct surveys among the general population (Groves 2005). This is mainly because of the fact that response rates have been slowly decreasing over the past decades (de Leeuw & de Heer 2002). Although this does not necessarily mean that nonresponse bias have been increasing as well (see Groves & Peytcheva 2008) survey researchers are nowadays trying to tailor survey designs to limit survey costs, keep up response rates and limit nonresponse bias. One of the ways in which surveys are tailored is by implementing mixed-mode survey designs. This paper discusses how to study one of the possible downsides of mixed-mode surveys: the mode effect. A mode effect occurs when respondents give different answers solely because of the method of interviewing. Studying mode-effects is difficult, because they are easily confounded with selection effects that occur when conducting surveys with multiple modes. This paper proposes Propensity Score Matching (PSM) as a method to disentangle mode effects from sample composition differences and shows how mode effects occur in a mixed Internet-telephone study.

3.1 Mode effects in mixed-mode surveys

In mixed-mode surveys, two or more methods of survey data collection are combined. The most prominent modes in current survey research are face-to-face, telephone, paper and the Internet (de Leeuw 2005). These modes can be combined in different stages of the survey process: to contact people, in the initial response phase, and also in following up on respondents.

While mixed-mode surveys intend to reduce potential coverage and nonresponse bias, this advantage may be offset by the occurrence of a mode-effect. A mode-effect occurs when respondents answer differently to a survey question, solely because of the mode in which the question is being administered. Mode effects might stem from differences in question administration: whether an interviewer is present, the media in which questions are administered and the way in which information is transmitted (de Leeuw 2005). These differences have led survey designers to worry about three related types of mode effects.

1 This chapter was co-written with Gerty Lensvelt-Mulders, Remco Frerichs and Assyn Greven and was published in the International Journal of Market Research, 53(5)

In situations where there is an interviewer, some people adjust their answers to what they expect the interviewer wants to hear. This social-desirability effect increases with the sensitivity of the question (Kreuter, Presser, & Tourangeau 2008). This leads to generally more positive answers when respondents evaluate a question on a negative-positive dimension, as will be the case in this study.

The second type of mode-effect can occur because of a difference in auditive versus visual transmission of data. In telephone surveys, interviewers read out the survey-questions along with all possible answer categories. The respondent listens and typically awaits the interviewer's instructions before answering. Those answer categories that are read out last, are more likely to be memorized and chosen (recency effect). In contrast to this, respondents in mail or Internet surveys read the questions and answer categories themselves. They read top-down or left-right and pick the first answer category that is thought to be appropriate (primacy effect) (Dillman & Christian 2005)

Finally, another mode-effect can occur with the choice for a "don't know" response category in telephone and Internet surveys. In telephone surveys, this option is generally not offered to respondents, but can be registered by the interviewer when respondents have trouble answering a question. In an Internet-survey the "don't know" option however is either explicitly offered or not offered, leading to differences in the frequency of "don't know" answers in a mixed-mode survey (Dillman & Christian 2005).

Although worries about mode effects have been extensively discussed in the survey literature, there is mixed evidence for their existence (for an overview, see de Leeuw 2005). Partially, this may be due to the fact that mode effects depend on the topic and specific structure of the question and response scale (Dillman et al. 2009). It also depends however on the fact that mixed-mode surveys lead to different compositions of the sub-samples. A difference that is found between two samples in a mixed-mode survey might be due to different levels of nonresponse or coverage bias in the different survey modes, but it could also be caused by a mode-effect.

3.2 Separating mode–effects from differences in sample composition

There are a number of ways to separate sample composition effects from mode-effects. Every approach has its disadvantages, and it is generally difficult to separate the two effects. The first and most straightforward way to assess nonresponse bias and mode-effects uses an experimental setting, in which a random group of respondents changes survey modes

during the interview (Heerwegh 2009). It is essential in such a design that none of the respondents who have to switch, drop out, and so this approach is difficult to use in a study among the general population.

A second approach is the comparison of survey estimates from mixed-mode studies to a 'golden' standard (de Leeuw 2005; Kreuter, Presser & Tourangeau 2008). One of the problems is that we seldom have validation data on attitudinal questions, which is the type of question where survey researchers worry about mode effects.

The third approach relies on statistical modeling. The goal of this approach is to make the two samples from a mixed-mode study equivalent. This can be done by weighting (Lee 2006), or by using a multivariate model that corrects for differences between the samples (Dillman et al. 2009). Finally, Latent Variable models can be used in combination with re-interviewing (Biemer 2001) or validation data on voter turnout (Voogt & Saris 2005) to correct for nonresponse bias. The disadvantage of these modeling approaches is that they assume that every survey mode can potentially cover the entire population. We know however that for example telephone and Internet coverage rates are not universal (Blyth 2008).

This paper takes a different approach and will show how Propensity Score Matching (PSM) can be used to match respondents from two sub-samples in a mixed-mode survey and study mode-effects. The idea of PSM stems from quasi-experimental research, and is used to eliminate differences in sample composition using a set of covariates. In this paper, we use PSM to correct for sample differences in levels of coverage and nonresponse. An illustration of this idea for the Netherlands is shown in Figure 1.

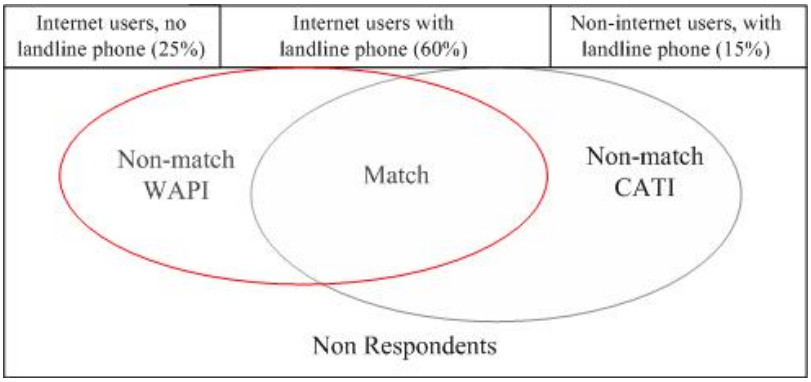

Figure 1: A mixed-mode survey where respondents from sub-samples are matched.

Notes: The sub-samples comprise three strata within the population:
1. Internet users without landline-phone (not covered by CATI),
2. non-Internet users with a landline phone (not covered by WAPI) and
3. those people covered by both Internet and telephone. The strata represent the coverage rates of landline phones and Internet in the Netherlands as of 2009 (Kool et al. 2009).

As opposed to weighting, PSM does not try to make the two samples equivalent. In fact, one of the main advantages of PSM is the fact that we can identify those respondents who are unique to a specific survey mode, and those who are found in both modes.

Matched respondents from the two survey modes share the same background characteristics. We argue that after matching they then should also be similar on other aspects related to the variables used in the matching process. Substantive differences that we find after matching for the matched respondents should be small, if there are no mode effects. If large differences remain after matching, they are likely due to mode-effects.

In the next section, we explain how we use three samples in this study: we first compare a probability-based Internet-sample to a quota sample drawn from an Internet-panel. We choose to first compare two Internet-samples in order to show how propensity score matching can be used to explain differences due to coverage and nonresponse bias between two samples. As all respondents in these two modes receive the same Internet questionnaire, mode effects cannot exist. We will show that differences between the matched Internet samples disappear after matching. In the second part we return to the primary objective of this paper, and match a telephone sample to the probability-based Internet-sample. We expect a mode effect after matching: the telephone respondents should respond more positive to a set of rating scales than the Internet respondents due to the presence of an Interviewer. Second, we expect the matched

samples to differ in the proportion of extreme positive as well as negative answers due to a recency effect in the telephone sample.

3.3 Methods

Sample

Our data stem from a mixed-mode survey conducted between April and June 2008 in the province of Zuid-Holland in the Netherlands. In the survey, respondents were asked how they experience environmental pollution from industry, traffic and agriculture.

For the survey two random samples were drawn from the central database containing all postal addresses in the Netherlands. The Computer Assisted Telephone Interview (CATI) sample consists of 6118 households which have a known landline phone number. They received a letter sent by the province government. A week later these households were called and the household member with the next birthday was asked for the telephone interview, with no incentive offered. Five recall attempts were made, if no contact was established. This procedure resulted in 2685 complete CATI-interviews and a response rate (RR1) of 47 per cent (AAPOR 2008).

The Internet-sample was also drawn from the central address-database. Because we lack a sampling frame of e-mail addresses for the general population we used a two step approach. 7090 households were sent a letter, which included an URL and an individualized login code to complete the survey on the Internet. Two weeks later, nonrespondents to the letter were sent a reminder by mail, and again two weeks later, nonresponding households who had a known telephone number, were phoned and asked to participate. Among those who participated in this Web Assisted Personal Interview (WAPI)[2] hundred gift vouchers each worth fifty Euros were raffled. This resulted in 1347 complete interviews and a response rate (RR1) of nineteen per cent (AAPOR 2008).

In order to investigate nonresponse bias and mode-effects, we drew a third sample in addition to the two probability-based samples. A quota sample stratified on age, gender and employment situation was drawn from the TeamVier access panel. Five hundred respondents took part in the exact same Internet-survey as the WAPI-respondents.

[2] It is more common to refer to Computer Assisted Self Interviewing over the Internet as Web Interviewing or CASI. In this dissertation I use both CASI and WAPI to refer to Web Interviewing, to be sure that terminology was consistent with the journal articles derived from the individual chapters

Instruments

The questions of the CATI and WAPI surveys were identical, except for the introduction and end of the questionnaire. Both surveys contained socio-demographic questions, including age, gender, highest level of education (7 point scale), composition of the household and employment status. From the postal code provided by the participants, we coded the degree of urbanization and average income in the street of the respondent on the basis of the registry of Statistics Netherlands (2009).

A set of seven questions asked how respondents experience environmental hindrance; our dependent variables. Respondents had to indicate on a scale from 1 (a lot of hindrance) to 10 (no hindrance at all), how much hindrance they generally perceived. The items asked for hindrance in the form of 1) dust from industry 2) bad smell from industry 3) noise from industry 4) bad smell from traffic 5) noise from traffic 6) noise from airplanes and 7) light pollution. A 'don't know' option was implicitly offered, both in the CATI and Internet-survey, where respondents could skip a question. We will assess mode-effects for all seven variables separately by evaluating the response patterns for all these variables in detail. We will also look at the combined composite score of the seven environmental hindrance questions to see whether mode-effects are consistent across variables, or cancel each other out[3].

Propensity score matching

Originally, propensity score matching models were developed to solve a problem in quasi-experiments. Individuals cannot always be assigned randomly to a treatment or control condition, as a result of which the estimation of treatment effects may be biased (Cook, Shadish, & Wong 2008; Deheji & Wahba 2002). This problem is similar to the situation in a mixed-mode study, where random assignment to one survey mode is in practice not possible because a respondent might not be able to respond in a specific survey mode (Schonlau, van Soest, Kapteyn, & Couper 2009).

The propensity score in our study summarizes the conditional probability to be a respondent in the CATI-sample, the WAPI-sample or the panel-sample. The propensity score indicates the differences between these samples pair wise. This means we compute three propensity scores of which two are of interest: first, the propensity to be a member of either

3 A Confirmatory Factor Analysis using Amos 7.0 (2006) yielded factor loadings for the composite score between .52 -.80, Cronbach's α .82. We computed a weighted mean score and use that variable as a composite score.

the CATI or WAPI sample, and second the propensity to be a member of the WAPI or panel-sample. We do not compare the CATI-sample to the panel sample. After propensity scores are computed, similar respondents from the WAPI and panel samples are matched based on their propensity scores which summarize their socio-economic background. Similarly, the CATI and WAPI respondents are also matched.

3.4 Results

Composition of the samples before matching

As expected, inability to participate and nonresponse in the CATI- and WAPI-samples lead to different coverage and nonresponse biases. Table 1 shows that the composition of the CATI and WAPI-samples differs significantly from the population before matching. The CATI respondents are older, are less often employed and single, are more often female, live more in non-urban areas and have a slightly lower monthly income than the general population. These results are in line with other nonresponse analyses of CATI-surveys (de Leeuw & van der Zouwen 1989). The only variable for which we somewhat surprisingly find no bias is level of education.

Table 1: Means and standard deviations for the socio-demographic characteristics of the respondents in the CATI, WAPI and panel-samples and the population

Independent Variables	Means (sd) CATI	Means (sd) WAPI	Means (sd) Panel	Population
Age	**55.1* (16.1)**	***50.1* (14.8)***	*45.9* (14.1)*	47.2
Employed (1=employed)	**.55* (.50)**	**.66 (.48)**	.67 (.47)	.67
Single (1=single)	**.31* (.46)**	**.22* (.42)**	.24* (.42)	.38
Gender (1=female)	**.56*** (.50)	***.45**** (.50)	*.51* (.50)	.51
Education (1-7)	**4.28 (1.73)**	**4.70* (1.54)**	4.85* (1.52)	4.24
Urbanicity (1-5)	2.35* (1.19)	2.44* (1.21)	-	2.22
Monthly net income (600-10000)	**2142*** (673)	**2287*** (687)	-	2200
Worries about society	**2.75** (.68)	**2.49** (.62)	2.53 (.63)	-
Knows environmental complaints agency (1=yes)	**.44** (.50)	***.39*** (.49)	*.56* (.50)	-
Dependent variables (1=a lot of hindrance, 10 – no hindrance at all)				
1) industry dust	**7.97 (2.45)**	***6.98 (2.63)***	*8.19 (2.39)*	-
2) bad smell industry	**8.05 (2.35)**	***6.98 (2.61)***	*8.42 (2.28)*	-
3) noise industry	**8.70 (2.07)**	***7.73 (2.51)***	*8.58 (2.30)*	-
4) traffic bad smell	**7.97 (2.38)**	***7.24 (2.50)***	*7.52 (2.54)*	-
5) traffic noise	**7.46 (2.66)**	***6.56 (2.72)***	*6.94 (2.64)*	-
6) airplanes noise	**8.52 (2.11)**	**7.80 (2.52)**	7.91 (2.40)	-
7) light pollution	**8.87 (2.04**	**8.18 (2.47)**	8.24 (2.49)	-
Composite score 7 items	**8.22 (1.58)**	***7.35 (1.82)***	*7.95 (1.84)*	-

Notes: * significant difference from population statistic with p=0.05 (one-sample t-test)
sd: standard deviation
Statistics in **bold:** significant difference between the CATI and WAPI-samples with *p*=0.05 (independent samples t-test)
Statistics in *Italics:* significant difference between the WAPI and panel samples with *p*=0.05 (independent samples t-test)
Population statistics are obtained from Statistics Netherlands (Statistics-Netherlands 2009)

The WAPI-sample is also biased. There is a significant difference for six of the seven demographic variables we tested. The only estimate that is in line with the population value is the proportion of people who is employed. For five variables (gender, household situation, education, urbanicity and income) the CATI-sample produces a less biased population estimate than the WAPI-sample, while the WAPI-sample is less biased on age and employment situation. For two variables (gender and income), the combined CATI and WAPI surveys would produce a good estimate for the population values, but for the other five variables, substantial biases would remain.

The WAPI and CATI-samples also differ on our dependent variables. Respondents in the CATI-sample consistently score significantly higher on all seven environmental hindrance questions. The differences are large. Table 1 shows that the means for CATI respondents are about 1 full point

or 10% higher than the means in the WAPI-sample. There are also differences between the WAPI and panel sample, although these differences are somewhat smaller. The question we now turn to is whether these differences in our dependent variables are caused by differences in sample composition or mode-effects caused by the different interviewing strategies.

Results from propensity score matching

As the propensity score is computed using a set of covariates, the choice of covariates is extremely important. We chose to use a basic set of socio-economic characteristics to compute the propensity score for ach individual: gender, being employed (dummy), age, household composition (single or not), education (1-7 scale), urbanicity (1-5 scale), income and knowledge of an environmental complaints phone number. We also use all possible two- and three-way interactions between these variables in a logistic regression analysis and compute a propensity score for every individual. These covariates produce a Nagelkerke R^2 of 0.16 in a logistic regression with survey mode (CATI-WAPI) as dependent variable. With WAPI-panel as dependent, we find a Nagelkerke R^2 of 0.17[4]. The socio-demographic variables produce a R^2 of 0.31 when the composite score serves as the dependent variable in a regression analysis. All these coefficients indicate that at least part of the nonresponse biases in the different samples can be explained by our covariates. The inclusion of more covariates would possibly increase the probability to explain all differences (Cook et al. 2008). The reasons why we constrain ourselves to this set of covariates are threefold. First, socio-economic variables are routinely used in marketing and social sciences to weight data. Second, socio-economic variables are highly correlated with access to both the Internet and a landline-phone. Finally, attitudinal variables are themselves subject to possible mode-effects, and therefore, we deem them unsuitable as covariates in this analysis.

Propensity score matching is implemented in the statistical programme R 2.9.1 (R Core Development Team 2009) along with the package 'MatchIt' (Ho, Stuart, Imai, & King 2009). Apart from being flexible, open-source and user-friendly, the 'Matchit' package offers many different ways to match respondents. We chose to use the technique of Coarsened Exact Matching (CEM) for two reasons. First, with CEM, the balance between the treatment groups is defined ex ante. This prevents the user

[4] Due to the fact that we do not have up-to-date information on income and urbanicity for the panel members, these values are not shown in table 2, nor were these variables used in matching the panel respondents to WAPI-respondents.

from adjusting imbalances through repeatedly running the matching procedure with different specifications for average treatment effect estimation error and number of matches. Second, CEM can deal with missing data, by discarding those cases from the matching procedure (Iacus, King, & Porro 2009)[5]. As a result, about 5 per cent of all respondents were not included in the matching procedure[6].

About sixty per cent of the Dutch population has access to both a landline phone and the Internet (Kool, Maris, & Munck 2009). For this reason we chose to match about sixty per cent of the sample members in our smallest sample (WAPI). For comparison reasons we specified about the same number of matches in the panel-sample[7]. Those respondents that were matched were as expected very similar on the covariates, leading to a balance improvement of 99 per cent. In other words, we managed to match about sixty per cent of the respondents in the panel and WAPI-sample to a very similar respondent in the WAPI and CATI-sample.

The WAPI and panel samples after matching

After matching, the WAPI and panel-respondents are according to our expectations very similar. From the WAPI-sample 209 respondents are matched to 162 respondents from the panel-sample. Most of the differences in the dependent variables that we found before matching disappear for these matched respondents.

[5] To make sure our results were robust, we also tried 'nearest neighbour', 'exact' and 'genetic' matching and in each of these methods we used various matching- specifications. In most settings, we arrived at the same results, although some settings did produce different results from the results we present here. We come back to this point in the discussion section.

[6] The R-code used for the matching procedures is available from the authors upon request

[7] Due to a smaller number of cases in the Internet-sample and greater imbalance in the propensity score, 45 per cent of all respondents in the panel sample was matched.

Table 2 shows the response patterns of both the matched and unmatched respondents. For the first of the seven questions on environmental hindrance, we see that the significant difference that we found before matching (as shown in table 1) is greatly reduced. Before matching, the mean hindrance score in the WAPI-sample was 6.98 and in the panel-sample 8.19. After matching, the hindrance for the matched WAPI-respondents is 7.58 and 8.07 for the panel-respondents. This difference is no longer significant. For the means of the other environmental hindrance questions, we find that the differences that were there before matching are consistently reduced after matching. The only strong difference that remains is for the question about the bad smell of industry. Two other differences remain marginally significant, while the differences on the other questions, as well as the composite score, disappear after matching (see Appendix A for all statistical tests).

Table 2: Differences between WAPI and panel sample after matching

		Mean	Sd.	% Pos.	% Extr. Pos.	% Extr. Neg.	% DK	N
Dust Industry	Match-WAPI	7.58	2.26	79.3	24.0	1.0	.5	208
	Match-panel	8.07	2.47	84.0	39.5	3.7	1.7	162
	Nmatch-WAPI	6.86	2.68	68.5	21.4	3.3	2.6	1082
	Nmatch-panel	8.24	2.36	85.7	44.3	2.7	4.0	300
Bad smell industry	Match-WAPI	7.35	2.41	77.5	22.5	3.3	0.0	209
	Match-panel	8.27	2.34	85.2	43.8	3.1	1.7	162
	Nmatch-WAPI	6.91	2.64	69.2	20.7	4.2	1.1	1099
	Nmatch-panel	8.48	2.28	88.9	48.9	3.6	1.9	307
Noise Industry	Match-WAPI	8.13	2.31	85.2	34.9	2.9	0.0	209
	Match-panel	8.43	2.35	87.8	45.7	3.7	0.6	164
	Nmatch-WAPI	7.65	2.54	79.7	31.8	3.6	1.3	1096
	Nmatch-panel	8.61	2.32	88.6	55.2	2.9	2.2	306
Bad smell traffic	Match-WAPI	7.69	2.34	80.4	29.2	2.9	0.0	209
	Match-panel	7.47	2.39	78.5	20.2	2.5	1.2	163
	Nmatch-WAPI	7.14	2.53	73.5	20.4	3.5	1.1	1099
	Nmatch-panel	7.49	2.64	78.5	28.3	4.6	1.9	307
Noise traffic	Match-WAPI	7.23	2.40	76.0	16.8	3.8	0.5	208
	Match-panel	6.71	2.56	66.1	12.1	3.0	0.0	165
	Nmatch-WAPI	6.45	2.75	64.9	13.0	6.5	0.6	1099
	Nmatch-panel	7.00	2.69	69.9	21.0	3.9	1.2	309
Noise airplanes	Match-WAPI	7.80	2.66	80.4	34.9	5.3	0.0	209
	Match-panel	7.68	2.51	80.0	30.9	3.6	0.0	165
	Nmatch-WAPI	7.79	2.50	81.6	31.5	3.4	0.7	1103
	Nmatch-panel	8.01	2.37	82.6	36.1	2.6	0.9	310
Light pollution	Match-WAPI	8.18	2.57	82.8	42.6	3.8	0.0	209
	Match-panel	8.23	2.33	86.1	43.0	2.4	0.0	165
	Nmatch-WAPI	8.17	2.46	85.1	41.0	3.3	1.1	1099
	Nmatch-panel	8.19	2.61	85.1	45.8	4.2	1.5	308
Composite Score 7 items	Match-WAPI	7.71	1.73	89.5	-	-	-	209
	Match-panel	7.82	1.78	82.7	-	-	-	165
	Nmatch-WAPI	7.27	1.83	88.1	-	-	-	1111
	Nmatch-panel	7.97	1.89	91.4	-	-	-	309

Notes: match-WAPI and match-panel refer to those groups of respondents that could be matched to each other. Nmatch-WAPI and nmatch-panel refer to the groups of respondents that were not matched.

Summary of findings: Nonresponse and coverage bias between WAPI and panel samples are explained

-no differences in means matches	panel 4x >WAPI, WAPI 3x > panel
-no recency effect (extreme positives) in matches	panel 4x >WAPI, WAPI 3x > panel
-no primacy effect (extreme negatives) in matches	panel 3x >WAPI, WAPI 4x > panel
-no acquiescence/social desirability in matches	panel 5x >WAPI, WAPI 2x > panel
-no differences in choices “don’t know”	too few cases to draw conclusions

Apart from the means, we also find the response patterns in the matched WAPI and panel-samples to be similar. There are differences in the proportions of positive responses within the matched samples, but neither the matched WAPI-, nor the panel-respondents are consistently

more positive (matched WAPI–matched panel differences range between -7.7 and +9.9 per cent).

In the proportion of extreme positive and negative responses we also find no consistent pattern for the seven dependent variables (differences in extreme positives range between -15.5 and +9.2 per cent and the difference for extreme negatives between -2.7 and +1.7 per cent). The response patterns of the matched WAPI and panel respondents are in conclusion very similar. The only indicator where differences persist after matching is the mean score on hindrance from bad smell from industry. We are not able to explain why a difference persists for this variable. All other indicator show that PSM is able to explain the differences caused by different levels nonresponse and coverage errors.

Apart from the matched respondents, Table 2 also shows the response patterns for the respondents that we were unable to match. In short, we find the unmatched panel-respondents to respond more positively in general, and choose the extreme positive answer category more often than the unmatched WAPI-respondents. As expected, the unmatched panel- and WAPI-samples do differ from each other.

Concluding, we find that propensity score matching successfully explains the differences between the matched WAPI and panel samples. Matching can be successfully used to select those respondents that are found in the two modes, as well as identify those respondents unique to a survey mode.

The CATI and WAPI-samples after matching

Matching the CATI and WAPI-samples proved to be more difficult than matching the two Internet-samples. Before matching, the means of all seven dependent variables as well as the composite score were different in the CATI and WAPI-samples. From table 3 we see that these differences are only slightly reduced by matching. The means for the 1068 respondents from the CATI-sample who are matched are still consistently higher than the means for the 708 WAPI respondents (see Appendix A for all statistical tests). This finding holds for all seven environmental hindrance questions as well as the composite score and indicates a mode-effect: respondents in the CATI-sample give consistently more positive answers than WAPI respondents, who are very similar to them. This is likely because of an interviewer effect.

Table 3: Differences between CATI and WAPI-sample after matching

		Mean	Sd.	% pos.	% Extr. Pos.	% Extr neg.	% DK	N
Dust Industry	Match-CATI	7.88	2.37	81.3	37.6	1.4	.9	1058
	Match-WAPI	7.27	2.43	75.7	22.2	1.9	2.1	688
	Nmatch-CATI	8.03	2.50	82.1	45.4	3.3	1.4	1534
	Nmatch-WAPI	6.63	2.85	64.0	21.7	4.2	2.2	595
Bad smell industry	Match-CATI	7.92	2.34	82.9	38.1	1.5	.6	1062
	Match-WAPI	7.28	2.45	75.9	23.0	2.6	.8	697
	Nmatch-CATI	8.15	2.35	84.8	45.9	1.8	1.1	1539
	NmatchWAPI	6.65	2.75	64.7	18.9	5.8	1.0	603
Noise Industry	Match-CATI	8.63	2.06	85.2	34.9	2.3	.6	1061
	Match-WAPI	7.90	2.38	82.2	33.2	2.7	1.0	696
	Nmatch-CATI	8.74	2.08	90.4	57.9	2.1	.4	1549
	Nmatch-WAPI	7.51	2.67	78.2	31.1	4.7	1.3	601
Bad smell traffic	Match-CATI	7.90	2.35	80.4	29.2	2.1	.6	1062
	Match-WAPI	7.45	2.35	78.2	21.8	2.7	1.0	696
	Nmatch-CATI	8.04	2.39	83.5	42.2	2.2	.5	1548
	Nmatch-WAPI	6.98	2.66	70.7	21.7	4.1	.8	604
Noise traffic	Match-CATI	7.34	2.59	75.6	28.4	3.2	.3	1065
	Match-WAPI	6.71	2.61	68.8	13.0	4.7	.7	698
	Nmatch-CATI	7.56	2.68	78.2	35.2	4.5	.4	1550
	Nmatch-WAPI	6.36	2.85	63.7	14.2	7.9	.5	606
Noise airplanes	Match-CATI	8.35	2.16	87.6	43.6	1.3	.2	1066
	Match-WAPI	7.93	2.33	83.7	31.5	2.5	.6	699
	Nmatch-CATI	8.65	2.06	90.8	52.7	1.5	.3	1551
	Nmatch-WAPI	7.64	2.70	79.0	32.7	4.8	.5	605
Light pollution	Match-CATI	8.78	2.10	90.5	58.1	2.2	.4	1064
	Match-WAPI	8.35	2.29	87.2	41.9	2.9	.8	697
	Nmatch-CATI	8.94	1.98	92.4	62.0	2.1	.3	1552
	Nmatch-WAPI	7.98	2.66	82.1	40.8	3.3	1.0	603
Composite score 7 items	Match-CATI	8.11	1.57	94.9	-	-	-	1068
	Match-WAPI	7.54	1.69	91.2	-	-	-	703
	Nmatch-CATI	8.30	1.59	95.2	-	-	-	1556
	Nmatch-WAPI	7.10	1.95	84.9	-	-	-	609

Notes: match-CATI and match-WAPI refer to those groups of respondents that could be matched to each other. Nmatch-CATI and nmatch-WAPI refer to the groups of respondents that were not matched.

Summary of findings: Differences between CATI and WAPI-samples remain after matching: occurrence of mode-effects.

-differences in means matches	CATI 7x >WAPI
-recency effect (extreme positives) in matches	CATI 6x >WAPI, WAPI 1x > CATI
-primacy effect (extreme negatives)in matches	WAPI 7x >CATI
-acquiescence/social desirability in matches	CATI 7x >WAPI
-no differences in choices “don’t know”	too few cases to draw conclusions

Unsurprisingly, the higher mean scores in the CATI-sample are accompanied by other differences in the response patterns. We find that the differences in the proportion of positive answers are consistently higher in the matched CATI-sample (differences between CATI and WAPI proportion of positive answers range between +2.2 and +9.4 per cent). We also find the matched CATI respondents are much more likely than matched WAPI-respondents to choose the most extreme positive answer category (differences between +1.7 and +16.2 per cent). In the WAPI-sample, respondents choose the extremely negative answer category more often than CATI-respondents for six of the seven variables. The differences are however small (between +0.4 and +1.5 per cent).

All in all, we believe our findings indicate two related mode effects: respondents in the CATI-sample are more positive than respondents in the WAPI-sample, even after matching. They also pick the extremely positive answer category more often, but this may partially be explained by the fact that CATI-respondents are more positive in general.

The differences in the response patterns of the matched CATI and WAPI-samples are not caused by a failure to effectively match respondents. Table 4 shows that the differences in the response patterns for the unmatched samples are even more pronounced than the matched samples. The differences in means, the proportions of positive responses and extreme responses are all larger in the unmatched samples than the matched samples. In the next section we discuss the implications of these mode-effects.

3.5 Conclusion and discussion

When carefully utilized, mixed-mode surveys can both increase coverage- and nonresponse rates and decrease bias resulting thereof. However, using different survey modes results in a confounding of sample selection effects and mode-effects, and separating these effects from each other is difficult. The starting point of this paper was to show how propensity score matching can help to disentangle mode-effects from sample effects.

Propensity score matching can be used to classify respondents who are unique to a certain mode versus respondents who are present in both modes. When two Internet-samples (panel-WAPI) are compared, the *matched* respondents from both samples are similar not only on their socio-economic characteristics; after matching they also show similar answer patterns on our outcome variables. This leads to the conclusion that propensity matching explains differences caused by sample selection effects. As expected, we find no mode-effects comparing the random

WAPI-sample and quota sample from an access panel. The differences in outcomes of the unmatched parts of these Internet-samples are due to differences in the compositions of the unmatched samples.

However, when *matched* respondents of the telephone and Internet-sample are compared (CATI-WAPI), respondents that appear to be similar on their background characteristics, still respond differently. Although the magnitude of the differences declines for the matched samples, the answer patterns of the matched samples show mode-effects. The matched CATI-respondents choose the extremely positive category more often and respond more positively in general than their WAPI-counterparts.

Concluding, we showed that mode-effects and nonresponse effects interact in mixed-mode surveys combining telephone and Internet surveys, making it impossible to straightforwardly merge the data from these surveys and analyze them as one dataset.

A limitation of our study is the way in which we studied mode-effects. The different mode effects that we wanted to distinguish (i.e. recency effects, primacy effects and interviewer effects) interact with each other, making it impossible to evaluate which types of mode-effects occur. Recency effects and social desirability in telephone surveys both lead to higher sample means, and in our study, it is impossible to separate the two.

A second limitation of our study is that propensity score matching is a form of statistical modeling related to regression techniques. As such, it suffers from some of the weaknesses that statistical models in general suffer from. A different specification or the inclusion of different covariates could have resulted in different results. We tried various matching specifications, and as long as we chose not to match all sample respondents, our results were robust. However, more research is needed on propensity score matching and its effectiveness in mixed-mode surveys to learn about the differential effects of matching specifications under different circumstances.

Looking forward, the central question that emerges in mixed-mode survey research is whether we can combine data from mixed-mode surveys. Here we offer two directions for further study. The directions both involve the use of external validation data. Adding substantive questions (e.g. newspaper readership), for which the aggregate population estimate is known, can be used to evaluate the quality of mixed-mode samples before and after matching. Moreover, external validation can give insight in the possible trade-off between nonresponse error and mode effects, and ultimately it is the trade off between errors

of non-measurement and measurement that researchers need to understand.

The combination of two mixed-mode samples in presence of mode effects is an issue that still needs to be taken up. Simply combining the two surveys and ignoring mode-effects does not seem the most sophisticated solution. The first and best solution to this problem is to try and prevent mode-effects. Unimode-questionnaires try to make questions cognitively equivalent across modes, reducing the problem of mode-effects (Dillman & Christian 2005).

A second method would be to assess mode effects first, and then decide whether the results from two modes should be presented separately or not. Propensity score matching can disentangle mode-effects from sample differences and shed light on this issue.

4 Chapter 4: Evaluating the effect of Dependent Interviewing on the quality of measures of change[8]

Panel surveys collect data from the same individuals over time in order to measure change and stability at the individual and macro level. Apart from the types of measurement errors that are also present in cross-sectional surveys, panel surveys suffer from specific forms of longitudinal measurement error that jeopardize data quality. The topic of this paper is longitudinal measurement error caused by spurious change. Estimates of change between two waves of data collection (from here on called wave) are often biased; estimates are either too small or too large when compared to the true value (Groves 1989). Earlier studies have shown that estimates of change in income are on average more likely to be overestimated than underestimated (Hoogendoorn 2004; Lynn, Jäckle, Jenkins, & Sala 2006; Webber 1994).

We argue in this paper that this is caused by measurement error in the variable of interest. Measurement error can occur when respondents form an answer (for example misinterpreting the question, forgetting, or a wrong estimation strategy), or when they misreport an answer (by coding errors or mistyping). When respondents experience the exact same difficulties in every wave – or put it differently – when the error made by each respondent is consistent across waves, change estimates will be unbiased. Most of the measurement error across waves however, is not systematic but random, leading to different sizes and directions of error at different waves and overestimations of estimates of change at the individual level: spurious change.

One solution that directly tackles the problem of spurious change is Dependent Interviewing (DI). DI is the process of providing respondent answers from a previous wave of the study to the current wave. In variations of Proactive Dependent Interviewing (PDI), respondents are presented their answer from the previous wave in the survey question. In the 'remind, continue' PDI-design, respondents are reminded of their earlier answer, but are then asked the normal (independent) survey question. In the 'remind, still' design, respondents are asked whether their old status is still the same, while in the 'remind, change' design the opposite question is asked: whether their status has changed (Jäckle 2009). As opposed to PDI, respondents in Reactive Dependent Interviewing (RDI) always have to answer the independent survey question first. Only when the data from the previous and current wave are inconsistent, or when a 'large' change occurs for quantitative

8 This chapter was co-written with Gerty Lensvelt-Mulders.

variables do they receive feedback on their answer from the previous wave, after which they can adapt or explain their answer. RDI greatly reduces coding errors (in interviewer administered interviews) and mistyping (in self administered), as any inconsistent answers between two waves will be directly fed back to the respondent (Jäckle 2009). The memory cue that PDI provides ideally alleviates the cognitive burden while respondents search for an answer, making it easier to complete the retrieval and judgment phase of the answering process (Tourangeau, Rips & Rasinski 2000). This is believed to lead to a reduction in measurement error in the variable of interest.

Both RDI and PDI have been successfully implemented in the British Household Panel Study (BHPS), the American Health and Retirement Study (HRS) the Survey of Income and Programme Participation (SIPP) and several other panel surveys. In these studies, DI has mainly been used to measure qualitative variables: employment status, family situation and income sources of the respondent. (Callegaro 2008; Jäckle & Lynn 2007; Jäckle, Laurie, & Uhrig 2007; Lynn, Buck, Burton, Jäckle, & Laurie 2005; Lynn et al. 2006; Mathiowetz & McGonagle 2000).

In this paper, we focus on the use of DI in questions about income amounts. The Canadian Survey of Labour and Income Dynamics (SLID), SIPP and BHPS have included DI procedures for income amounts from labour, assets and government transfers (Dibbs, Hale, Loverock, & Michaud 1995; Hale & Michaud 1995; Hill 1994; Jäckle et al. 2007; Lynn et al. 2006; Moore, Bates, Pascale, & Okon 2009; Moore, Bates, Pascale, Griffiths, & Okon 2006; Webber 1994). In most panel surveys, personal and household incomes are constructed from multiple questions. First, respondents indicate what types of income they receive. Follow-up questions typically ask about durations of receipt, amounts, and then depending on the panel survey, ask more details of receipt. The BHPS uses RDI to make sure the list of the income sources received is consistent with previous years (Jäckle et al. 2007). In the SLID, RDI is used throughout the income section, for example to record wage income amounts (Hale & Michaud 1995). Finally, in the SIPP, 'PDI – remind continue' is used for receipt of income sources, while income amounts are asked with PDI in case of initial item-nonresponse (Moore 2006).

Despite its widespread use, it is unclear what effect DI has on data accuracy, and data quality. Multiple authors (Bates & Okon 2003; Conrad, Rips, & Fricker 2009; Hale & Michaud 1995; Hill 1994; Jäckle 2009) have voiced concerns about the effect of DI on data quality, particularly with PDI. Any measurement error that is present in the data from wave t-1 could be fed forward to the current interview, possibly leading to correlated measurement errors across waves.

The cause for such correlated measurement error lies in the cognitive process of responding to a survey question. When respondents skip one or more of the steps in the response process and give an answer without thinking it through this is called satisficing (Tourangeau, Rips & Rasinksi 2000; Krosnick 1991). Satisficing in the context of Dependent Interviewing may either lead to falsely confirming the income-amount from the previous wave, or using incorrect data from the previous wave as an anchor for giving an answer in the current wave.

Hill (1994) notes that it is probably safe to assume that the positive effects of DI on data quality outweigh the potential negative effects by correlated measurement errors. Hoogendoorn (2002; 2004) and Holmberg (2004) both found that there were no more respondents reporting no change in a PDI-condition than in independent interviewing (INDI), signaling that satisficing might not be problematic. Conrad, Rips and Fricker (2009) find a similar result when using PDI in combination with questions where respondents had to recall dollar amounts from earlier surveys. They find that overestimates of change between two consecutive waves are reduced by PDI. However, when they look at the effect of PDI on measurement error in individual responses by validating them against the true dollar amounts, they find no improvement when the PDI-condition is compared to a control-condition. They argue that this is because the dependent facts that were provided sometimes contain measurement error themselves, and that the positive and negative effects of PDI do outweigh each other.

Overall, it is unclear if DI leads to any negative effects on data accuracy, whether any negative effects are outweighed by the positive effects of reducing spurious change, and how these effects behave under different levels of measurement error in the variable of interest. The goal of this paper is to experimentally test the effect of Dependent Interviewing on the data quality of income questions in a self-administered Web Interview. In addition, we try to experimentally manipulate the amount of measurement error by varying the amount of cognitive support that we include in the income question. This tests how the effect of DI behaves under different levels of measurement error. We have two positive and two negative expectations about the effect of DI on data accuracy. We expect:

1. Income questions that provide more cognitive support to lead to less measurement error,
2. DI to lead to less measurement error because of a reduction of reporting errors (RDI), and recall errors (PDI) in comparison to a control group,

3. RDI and PDI to cause correlated measurement errors because of satisficing. This effect is greater for PDI than RDI, because the risk of satisficing is much greater with PDI than RDI, and
4. Correlated measurement errors in RDI and PDI to be larger when the survey question provides less cognitive support.

4.1 Methods

Sample

The data in this study were collected in a cohort panel survey among all first year psychology students at Utrecht University in the Netherlands. The panel study was established in September 2007 and aims to follow students throughout their studies. A total of 444 respondents were invited to the first wave of the panel survey. They were approached in the first three weeks of their degree course by e-mail. An individualized URL would lead them to a website where they could start with the first survey. A total of 298 respondents completed the first survey, which resulted in an initial response rate (RR1) of 69 per cent (AAPOR 2008). We believe this relatively low response rate was partly due to the fact we contacted respondents at their university e-mail address. We later learned that several students were not yet actively using this e-mail address at that time.

Our study only includes data from the first four waves of the panel survey. These waves were held at three-month intervals. Re-invitations were sent to non-respondents, but not to 'hard' refusals or those who dropped out of their degree course. The attrition in the four waves of our panel sums up to about thirty per cent and is not selective for any experimental group; attrition is equal across conditions. A complete overview of response rates can be found in Table 4.

Table 4*:* Invitations, complete responses and response rates for the panel

Response rates	Invitations sent	Complete Responses	Response Rate
Wave 1	444	298	69%
Wave 2	298	255	86%
Wave 3	255	219	73%
Wave 4	219	202	68%

Notes: The response rates for waves 2-4 are conditional on response in wave 1. Complete respondents exclude respondents who did participate in the wave, but had item-nonresponse on the income question, and respondents who failed to participate in one or more of the four waves. The response rate (RR1 for wave 1, RR2 for waves 2-4) was computed according to the definitions of the American Association for Public Opinion Research (AAPOR 2008)

Design

DI was used to measure various constructs. In this paper we focus on the use of DI for income questions for two reasons. First, income is an important and frequently measured construct in surveys, be it in government-sponsored surveys, academic research, or marketing. Secondly, income is likely to change in a first year student population, especially as we followed students at the start of their degree course, when they usually start to take on a student loan or a side job.

In this study we focus on general net income-amounts. We designed an experiment that tests the effects of DI under different levels of cognitive support using a 3 x 2 factorial design varying both the DI-procedure (control, PDI, RDI) and the extent of measurement error in the income question (vague, specific question). Students were randomly assigned to one of the following six conditions.

The first experimental condition served as a control condition in which respondents only receive the independent question. “what is your approximate monthly income?” The second experimental condition uses ‘remind continue’ PDI: “in our last survey, you reported to have a monthly income of approximately {OLD VALUE}. What is your income at this moment?” In the third experimental condition we use RDI. Respondents first answer the independent question (see control condition). When a wave-on-wave change in income larger than 10 per cent occurs, they receive a follow-up question: “In our last survey you reported to have a monthly income of approximately {OLD VALUE}. Did your income change?” with possible answer categories: “1. no, my income is still {OLD

VALUE}" and 2. "yes, my income changed to {FILL IN}, because of {FILL IN}". In our analyses, we used only the data after the follow-up questions.

In order to further gauge the effect of DI on data quality we also tried to manipulate the amount of measurement error in the income question. There are many ways in which the reliability of a question can be manipulated, as answer scales, question introductions and question layout can all affect the reliability of a survey question (Saris & Gallhofer 2007). We chose to focus on the specificity of the introduction of the income question and make a distinction between specific and vague descriptions of the income sources to consider in forming an answer. In the specific condition respondents received the question: "what is your approximate monthly income? Think about a student loan, income from your job, money from your parents and/or other sources", while in the other they received the vague question "what is your approximate monthly income?" preceded by a short introduction to make the two questions of equal length (Saris & Gallhofer 2007). As the specific question provides cues for the most common income sources to consider, it is less likely that respondents forget to include any income in some waves but not in others (Moore, Stinson, & Welniak 2000).

The vague and specific questions were introduced in the first wave of the study, while DI was introduced in the second wave. Respondents remained in the same experimental condition throughout the study, except when income data from the previous wave were missing. In that case they received the Independent question and were further excluded from our analyses. The 3x2 factorial design resulted in six experimental groups.

Modeling data quality – the quasi simplex model

As we do not have register data on income to validate our results, we use a statistical model that can separate the positive and negative effects of DI on data quality. The statistical model that we use is the quasi-simplex model as introduced by Heise (1969) and Wiley and Wiley (1970). The quasi-simplex model shares a common framework with other Structural Equation Models, that every observed score (Y_t) is composed of a latent true score (T_t) and some measurement error (E_t) for t = *wave 1,2,3...P* (see equation 1). The measurement error (E_t) has a mean of zero, and a normally distributed variance term (v_t) at every wave.

$$Y_t = T_t + E_t \qquad (1)$$

The second important assumption in the quasi-simplex model is that the true score at each wave (T_t) can be fully explained by the true score at the

previous wave (T_{t-1}), the stability between the true scores at t and t-1 ($\beta_{t,t-1}$) and a random disturbance term (ζ_t) which represents the true score change (or noise) over time. Parameter ζ_t follows a normal distribution with a mean of zero and variance term (Ψ_t).

$$T_t = \beta_{t,t-1} T_{t-1} + \zeta_t \qquad (2)$$

Equation 2 describes a Markovian process, in which the true and observed scores are only determined by the true scores at the earlier wave (see Alwin 2007; Saris & Andrews 1991 for an overview and thorough discussion of the variance decomposition in quasi-simplex models).

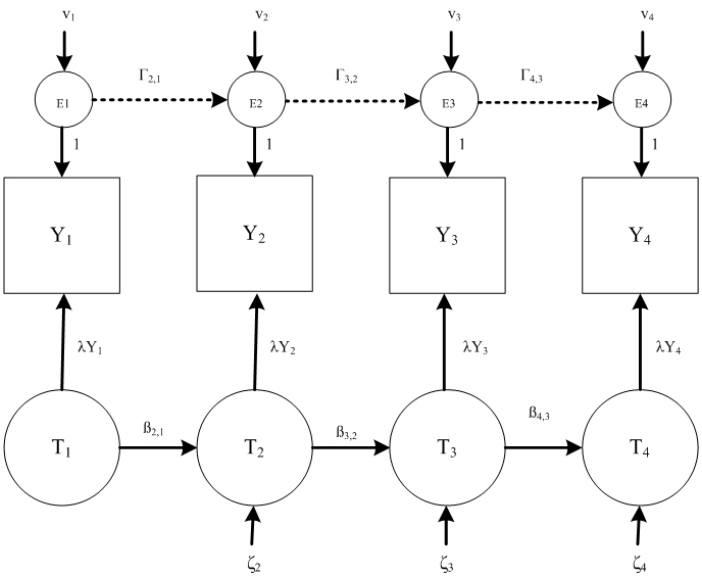

Figure 2: Basic quasi simplex model (only full arrows) and quasi simplex model with correlated measurement errors (full and dashed arrows).

The full arrows in Figure 2 depict the basic quasi-simplex model: the reliability of any survey question can be estimated, when either the loadings between the true and observed scores (λY_t) or measurement error variances (v_t) are assumed to be equal across waves. The assumptions lead to identical results, but as we will later assume the measurement errors to be related, we chose to constrain the factor loadings (λY_t) across all waves. The reliability coefficient (R) is derived by dividing the variances of the true scores ($V(T_t)$) by the variance of the observed scores ($V(Y_t)$), where the variance of the true score is computed as $V(Tt)$-E_t (see Alwin 2007).

The introduction of DI however makes it necessary to alter the quasi-simplex model and incorporate one further equation that allows for correlated measurement errors to mimic the effect that PDI and RDI might have on our data. Such models have earlier been proposed by Wiley and Wiley (1974) while Palmquist and Green (1992) showed that this model yields consistent estimates when data are available for four or more waves. The model with correlated errors is also depicted in Figure 2 and now includes correlated measurement errors $(\Gamma_{t,t-1})$, shown as the dashed arrows between the measurement errors (E_t).

Each observed score (Y_t) is now not only determined by the true score at the same wave, but also by the observed score and measurement error at *t-1*, leading to equation 3, being a revised version of equation 1:

$$Y_t = T_t + E_t + \beta_{t,t-1} T_{t-1} + \Gamma_{t,t-1} E_{t-1.} \quad (3)$$

Apart from the assumption that apply to the basic quasi simplex model, two more assumptions are now necessary for model identification; errors are only allowed to correlate between two consecutive waves, while the $\beta_{t,t-1}$-parameters need to be equal across all experimental groups (Palmquist & Green 1992). As we experimentally assigned people to experimental groups, it follows naturally that the stability parameters of the true scores should in fact be equal.

The resulting 'quasi-simplex model with correlated errors' allows us to disentangle the reliability (R) of a survey question, the true score stabilities $(\beta_{t,t-1})$ and change (ζ_t) between the consecutive waves and the extent to which measurement errors are fed forward in DI $(\Gamma_{t,t-1})$. The reliability coefficient (R) will give us insight in how much DI positively affects data accuracy by reducing spurious change, while the correlated measurement error (Γ) will show us how problematic satisficing is.

4.2 Results

We have complete data from all waves for 202 respondents. Four respondents were considered outliers, because they reported an income over 2500 euros a month (about 5 times the mean) and were excluded from the analysis. We used AMOS 16.0 to test the model (Arbuckle 2007). The distributions of the income questions are somewhat skewed and kurtose, but values for multivariate kurtosis (Mardia's index) and skewness remain below 3.0 (Kline 2005).

Descriptive statistics

Table 5 shows the descriptive statistics for the net income questions in the six experimental conditions. The income of our respondents gradually increases from a mean of 459 Euros in wave 1 to 578 in wave 4. There is a clear difference in means when we group the data for the first wave into the vague (conditions 1-3) and specific income questions (conditions 4-6) ($t(2,218)=2.57$, $p=0.01$). A repeated measures ANOVA including all four waves shows that the difference between vague and specific questions disappears in later waves ($F(2,217)=.35$, *n.s.*).

Table 5: Means and standard deviations of the reported income in the experimental conditions of the study

Means and standard errors (in parentheses) of monthly income in euros	Wave 1	Wave 2	Wave 3	Wave 4	Mean all waves
1) Vague, independent	441 (52)	533 (49)	554 (52)	572 (55)	522 (48)
2) Vague, PDI	409 (43)	419 (46)	498 (48)	547 (55)	449 (43)
3) Vague, RDI	375 (45)	431 (49)	487 (48)	536 (55)	458 (42)
4) Specific, independent	529 (57)	513 (64)	523 (74)	638 (105)	562 (55)
5) Specific, PDI	487 (41)	489 (42)	498 (48)	560 (57)	519 (42)
6) Specific, RDI	498 (47)	535 (61)	533 (58)	598 (69)	539 (50)
Mean within wave	459 (20)	488 (21)	520 (22)	578 (27)	512 (17)

Notes: N(total): 202 - condition 1: 34, condition 2: 38, condition 3: 37, condition 4: 30, condition 5: 37, condition 6: 26

Reported changes

Before turning to the statistical model of the effect of DI on data quality of estimates of change, we first show some of the observed change statistics. Table 6 shows that the proportion of students that report no changes in the PDI conditions (42%) is higher than in the other two conditions (RDI 34%, INDI 25%). Respondents in the control conditions logically report more small changes (>0% and <10%) than those in the PDI conditions (PDI 26%, RDI 29%, INDI 40%). The number of people reporting large changes > 10% and < 100%) is about equal in all groups.

Table 6 also shows that 39 per cent of all respondents in the RDI conditions receive feedback on their income reports, as their reports from consecutive waves change more than 10 per cent. About 10 percent of the people who receive RDI feedback subsequently change their answer. To contrast this finding, we also show the proportion of income reports where wave-on-wave change is larger than 100 per cent. Although the number of reports is here small, we see that these reports amount to between 4 and 16 percent of total reports, and that such outliers are relatively often (about 29% of the outliers) corrected with RDI.

Table 6: Proportions of change in income reports between two consecutive waves in all experimental conditions

Experimental conditions/ Proportions	Reporting no change – 0%	Reporting changes > 0% and <10%	Reporting changes > 10% and <100%	Reporting Changes >100%	Changing answer after RDI	Outliers corrected
1) Vague, independent	0.29	0.29	0.38	0.04	-	-
2) Vague, PDI	0.43	0.24	0.19	0.08	-	-
3) Vague ,RDI	0.26	0.29	0.29	0.16	0.12	0.25
4)Specific, independent	0.21	0.50	0.19	0.10		-
5) Specific , PDI	0.41	0.28	0.22	0.09		-
6) Specific, RDI	0.42	0.29	0.20	0.09	0.09	0.33

Notes: N(total) equals the total sample size multiplied by three wave-on-wave transitions: N(total): 606- condition 1: 102, condition 2: 114, condition 3: 111, condition 4: 90, condition 5: 111, condition 6: 78

The second, third and fourth column denote the proportion of respondents after DI with either no wave-on-wave change in income, or wave-on-wave change between 0 and 10 percent, and wave-on-wave change between 10 and 100 percent. We somewhat arbitrarily defined an income report as an 'outlier' (column 5) when the difference between t,t-1 before RDI-feedback was larger than 100 per cent.

The proportion of people that changed their answer after RDI (column 6) only includes those people that received feedback – hence reported an income change larger than 10%.

The proportion of corrected outliers (column 7) indicates what proportion of the people who reported an outlier (> 100% change) changed their answer after RDI.

Effect on data quality - model selection

We now move to the quasi simplex model to investigate the effect of DI on the data quality of change estimates. We first fit the basic quasi-simplex model shown as the model without dashed arrows in Figure 2.

This starting model (1a) as shown in Table 7 fits the data ($\chi 2(12) = 16.58$, p=.17, Root Mean Square Error of Approximation (RMSEA) =.042, Tucker-Lewis Index (TLI) =.97, Akaike's Information Criterion (AIC) = 160.58), according to rules of thumb for model fit (RMSEA <0.05 and TLI >0.95 (Kline 2005). From this point on we will compare any further restrictions that we impose on the quasi-simplex structure to this basic model (1a). As the more restricted models are always nested within the basic quasi-simplex model, we can use a measure of the difference in fit between the models ($\Delta\chi 2$), as well as the difference in the number of degrees of freedom (Δdf) to conduct a $\chi 2$-difference test with a p-value that indicates whether the more restricted model fits significantly worse than the less restricted model. As this measure is sensitive to overfitting in

models with a small sample size, we will use the values of AIC as a final model selection criterion (Bollen 1989; Haughton, Oud, & Jansen 1997).

The stabilities between the true scores for income should logically be the same across conditions, as we experimentally assigned respondents to a condition. The first restriction that results from this, is that the stabilities between the true scores (ßt,t-1) are equal across all six experimental conditions, leading to model 1b (see Table 7), in which we estimate 15 parameters less than in model 1a. This model 1b does not fit significantly worse than model 1a ($\Delta\chi2(15) = 13.17$, p=.59), and also has a better value for the AIC (143.75).

The second model we test is the quasi-simplex with correlated errors (the model with dashed arrows as shown in Figure 2). In the control conditions, the correlated measurement errors are set to 0. As this model fits the data well (see for fit measures model 2a in Table 4), we also add further restrictions to the model with correlated measurement errors, first by imposing the restriction that the correlations between measurement errors are equal for all t,t-1 (model 2b in Table 7) within each experimental group. This means that any effect of DI on the correlation of measurement errors is the same in all wave-on-wave transitions. As a final restriction, we constrain the measurement errors (E) to be equal at all t within all vague (1-3) and specific (4-6) conditions. This results in a model where the extent of measurement error (E) is the same for the three vague questions and three specific conditions (model 2c in Table 7), implying that DI does not affect measurement errors. This final, most restricted model, has the best fit to the data of all models looking at the value of AIC ($\chi2(29)= 33.50$, p=26, RSMEA=.027, TLI=.99, AIC=143.50). We tried to fit an even more restrictive model, in which all measurement errors (E) were constrained over all t and over all six experimental groups. If this model would fit, it would mean that the model is equivalent across all six experimental conditions, meaning that the experimental manipulation of measurement error by providing vague or specific cognitive support did not work. Although this model also fitted the data well, the AIC of this model is a bit higher than model 2c, meaning it fits relatively worse, and that our experimental manipulation of measurement error did work. As the AIC-value for model 2c is the best among the tested models, we will look at the parameters of this model to evaluate the effects of DI under different levels of measurement error on data quality.

Table 7: Results - the first six columns show the model fit statistics. The last three columns show the relative improvement of the model compared to the previous accepted model

Model/ test	χ^2	df	p-value	TLI	RMSEA	AIC	$\Delta\chi^2$	Δdf	p-value
1a)	16.58	12	.16	.97	.042	160.58	-	-	-
1b)	29.75	27	.33	.99	.022	143.75	13.17	15	.59
2a)	16.41	15	.36	.99	.021	154.41	-13.35	-12	.34
2b)	26.98	23	.26	.99	.028	148.98	10.58	8	.23
2c)	33.50	29	.26	.99	.027	143.50	6.52	6	.37
2d)	36.17	30	.20	.99	.031	144.17	2.68	1	.10

Notes: model constraints:
Basic quasi-simplex models:
1a) quasi simplex model
1b) model 1a + *ß* for every t,t-1 constrained to be equal between all 6 groups
Quasi-simplex models with correlated errors:
2a) model 1b + *Γt,t-1* freely estimated, set at 0 for independent interviewing
2b) model 2a + *Γ* for all t,t-1 equal within every experimental group
2c) model 2b + *E* equal for all t within vague and specific questions and *Γ* for all t,t-1 equal across the two PDI and two RDI experimental groups
2d) model 2c + *E* equal for all experimental groups

Parameter estimates for the quasi-simplex model with correlated errors

The final quasi-simplex model with correlated measurement errors includes constraints between all six experimental groups on the true score stabilities ($ß_{t,t-1}$) between waves, the correlations between measurement errors ($\Gamma_{t,t-1}$) and the measurement errors (E_t) within the three vague and three specific conditions. The stability ($ß_{t,t-1}$) between waves is large and consistent. The unstandardized coefficients are equal across all conditions, while the standardized coefficients range between .81 and 1.00 (see Table 8 and Table 9). As the stability between waves is high, the change (or instability) is low; coefficients for the disturbance of the true score (ζ_t) range between .01 and .34, showing that most of the variance in the true scores at wave t can indeed be explained by the model.

The incomes of students in our sample show a stable development over time; monthly incomes gradually increase over the waves. The amount of measurement error in the income measures in all six conditions is limited. The reliability coefficients (r) range between .93 and .99, indicating that even in the vague income questions, the reliability of income is sufficient. Reliability coefficients between .90 and 1.00 are quite common for income measures (Alwin 2007; Marquis, Marquis, & Polich 1986). The specific questions that include a list of the most common income sources are however more reliable than the vague questions as the model with equal reliabilities for all conditions did not fit the data as well as the model with unequal reliabilities.

The correlations between the measurement errors are small and do not reach significance. The standardized effects are about .08 for PDI and .02 for RDI indicating that in the RDI-conditions there is no risk of correlated errors, and that, if at all, this risk is small in the PDI conditions.

Table 8: Unstandardized coefficients in the six experimental conditions for the quasi simplex model with correlated measurement errors (model 2c)

Experimental condition/ parameter estimates	$ß_{2,1}$ (s.e.)	$ß_{3,2}$(s.e.)	$ß_{4,3}$(s.e.)	$Γ_{t,t-1}$(s.e.)
1)Vague, independent	.96 (.05)	.95 (.03)	.95 (.04)	-
2)Vague, PDI	96 (.05)	.95 (.03)	.95 (.04)	.08 (.06)
3)Vague, RDI	96 (.05)	.95 (.03)	.95 (.04)	.02 (.07)
4)Specific, independent	96 (.05)	.95 (.03)	.95 (.04)	-
5)Specific, PDI	96 (.05)	.95 (.03)	.95 (.04)	.08 (.06)
6) Specific, RDI	96 (.05)	.95 (.03)	.95 (.04)	.02 (.07)

Notes: all coefficients are significant at α<0.01, except for the correlated measurement errors ($Γ_{t,t-1}$) in all conditions. The factor loadings ($λY_t$) are not shown as they are constrained to 1 at all t in every group. The unstandardized coefficients for the unexplained variances of the latent variables ($ζ_t$) are also not shown, as we deem the standardized values of $ζ_t$ shown in table 6 to be more informative. Only one column of correlated measurement errors ($Γ_t$) coefficients is shown, as the coefficients are equal for every t,t-1.

Table 9: Standardized coefficients in the six experimental conditions for the quasi simplex model with correlated measurement errors (model 2c)

Experimental condition/ parameter estimates	R	$ß_{2,1}$	$ß_{3,2}$	$ß_{4,3}$	$Γ_{2,1}$	$Γ_{3,2}$	$Γ_{4,3}$	$λY_1$	$λY_2$	$λY_3$	$λY_4$	$ζ_2$	$ζ_3$	$ζ_4$
1)	.93	.96	1.00	.86	-	-	-	.97	.97	.96	.97	.07	.01	.25
2)	.94	1.00	.88	.97	.08	.07	.08	.97	.90	.91	.90	.01	.23	.05
3)	.94	.93	.87	.81	.02	.02	.02	.96	.95	.96	.97	.14	.25	.34
4)	.98	.90	.96	.87	-	-	-	.99	.99	.99	.99	.19	.09	.23
5)	.98	.91	.94	.89	.07	.08	.07	.99	.93	.92	.93	.17	.12	.21
6)	.97	.88	.96	.88	.02	.02	.02	.98	.97	.97	.97	.23	.07	.22

Notes: R= reliability coefficient. The reliability coefficient is computed as the true score variance ($V(T_t)$ divided by the observed score variance ($V(Y_t)$ (Alwin, 2007, Palmquist & Green 1992). All coefficients are significant at α<0.01, except for $Γ_{t,t-1}$ in all conditions. Explanation of the numbered experimental conditions can be found in Tables 2, 3 and 5.

4.3 Conclusion and discussion

Dependent Interviewing has been introduced in many panel surveys to reduce the extent of spurious change that occurs between two waves. Earlier studies showed that DI was successful at reducing such change, but there are worries that, especially in Proactive Dependent Interviewing, satisficing might lead to correlated measurement errors which negatively affect data quality. Our study showed that the negative

effect of correlated measurement error is nonexistent in RDI and, if at all, very small in PDI. On the other hand, there is also no positive effect of DI on data quality.

As we did not have validation data, we relied on statistical modeling to disentangle the positive and negative effects of DI on data quality. For this, we used a quasi-simplex model with correlated errors. Our results are robust, meaning that under different model specifications, we came to the same conclusion about the effect of PDI and RDI on data quality. The model with correlated measurement errors and additional constraints on some of the parameters however does the best job in showing how the effects of data quality are similar in our six experimental groups.

Our four expectations about the effects of DI on data quality were largely not confirmed. Our first expectation was met: we found that the vague income question contained more measurement error than the specific questions. However, we could not find evidence for our second expectation – that Dependent Interviewing (either DI or RDI) leads to less measurement error. We did find some evidence in accordance with our third and fourth expectations, that DI leads to correlated measurement error and especially so when the question provides less cognitive support. Both for RDI and PDI we found a non-significant effect. The sample size in our study was relatively small however, leading to low statistical power to detect any significant negative effects. It is likely that a larger sample size would lead to a significant negative effect of PDI in particular, although the effect size would remain small. A second limitation of our study is the fact that we used a student sample interviewed at three month-intervals. Most panel surveys interview respondents annually, and most respondents would have more stable incomes, possibly leading to different effects.

As first-year students undergo life changes that make the estimation of monthly income difficult for them, the fact that we find no large negative effect of either RDI or PDI on data quality is comforting. About 10 per cent of our respondents correct their incomes in the RDI conditions, indicating that DI is particularly effective for a small part of our sample. In the follow-up questions on the RDI-question, several respondents indicated that they mistyped their answer. The reduction of outliers or mistypes alone makes it worthwhile to implement DI. The disadvantage of the quasi-simplex model (and in fact most statistical models) is that they cannot point out who satisfices, and under which circumstances.

A few studies have tried to focus on explanations under what circumstances DI works and for whom. (Jäckle 2008a; Jäckle 2008b; Lynn et al. 2006). In our study we used the 'PDI, continue' design which should

be less prone to satisficing and acquiescence than either the 'PDI, remind' or 'PDI still' design. Future studies into DI should study who are likely to satisfice with PDI, and whether correlated measurement errors are larger in the 'PDI remind/still' design, which is currently used more often than the 'PDI, continue' design .

The use of Dependent Interviewing until now remains limited to factual variables, and most of them are qualitative in nature (for example occupation, industry codes and job status). The results from studies by Conrad, Rips and Fricker (2009) and Rips, Conrad and Fricker (2003) suggest that the way DI works to improve recall and reduce measurement errors are largely the same for qualitative and quantitative variables. Due to lower cognitive effort that qualitative variables mostly require from the respondent, the risk for satisficing might be lower for qualitative than for quantitative variables. This then suggests that PDI might be less problematic for qualitative than quantitative variables. An additional benefit of the use of RDI for quantitative variables is that it tackles the problem of mistyping, which is often only problematic for quantitative variables. Apart from this, the interviewing situation (self or interview-administered) might also affect the choice for a PDI or RDI design. More research on this issue is needed.

Finally, we note that it is important to decide when previous answers are fed forward in an RDI design with quantitative variables, and when they are not. Common sense suggests that feeding forward too many answers might lead to annoyance with the respondent, while feeding forward too few answers reduces the efficiency of DI.

We suggest that respondents see their previous answers only when changes occur that are too large to be caused by random measurement error. This first implies that it is very important to design valid and reliable survey questions. Then, the extent of unreliability or random measurement error can be used to determine when answers should be fed forward or not. We suggest that answers are only fed forward if changes between two waves are larger than can be expected based on the reliability coefficient.

5 Chapter 5: Can I just check…? Effects of edit check questions and Dependent Interviewing on measurement error and survey estimates[9]

Household income is a key measure of social welfare and as such important for policy analyses. Measuring household income is therefore one of the main purposes of many government-funded or other large-scale socio-economic surveys. Measuring household income is however difficult: it requires collecting information about all possible sources of income for each household member. Respondents need to remember all sources, as well as the timings of receipt and amounts received. This is a difficult and potentially tedious task. As a result, household income is likely to be measured with error, which may affect other derived estimates, such as poverty rates or income dynamics over time.

In this paper we assess the effects of edit check questions on estimates of household income and poverty status. We examine the effects of both within-wave and cross-wave edit checks. *Within-wave edit checks* use information collected earlier in the same interview to check the consistency of responses and detect potential reporting errors. Respondents are queried about sources they have not reported, but for which they are likely to be eligible, judging from responses given earlier in the interview (Pennell, 1993). *Cross-wave edit checks* are specific to longitudinal surveys. They use information provided in previous interviews to check the longitudinal consistency of responses. Respondents are queried about sources they have reported in the past, but not in the current interview (Jäckle, 2009; Mathiowetz & McGonagle, 2000) . In some surveys, such as the US Survey of Income and Program Participation, cross-wave edit checks are also used to verify whether apparent changes in amounts of receipt are genuine or due to a reporting or data entry error. Cross-wave edit checks are typically referred to as 'dependent interviewing' (DI) and we follow this convention. For simplicity, we refer to within-wave edit checks as 'edit checks'.

The key question examined here is to what extent DI and edit checks affect estimates of household income and poverty. Previous validation studies have shown that there is considerable under-reporting of non-labor income sources and that DI reduces this to some extent (Lynn et al 2012; Moore et al 2009). Although over-reporting also occurs, this is rare and does not change with DI. A couple of studies have examined the

9 This chapter was co-written with Annette Jäckle, and is due to appear as an ISER working paper in the summer of 2011. The second author gratefully acknowledges funding from the ESRC (RES-000-22-2323). Data collection for the experimental validation survey was funded by the ESRC Research Methods Programme (H333250031).

effects of DI on reported timing of receipt for individual income sources: Moore et al. (2009) showed that DI reduces biases in estimates of monthly transition rates and Jäckle (2008) showed that DI can improve estimates of spell durations, in particular reducing the under-reporting of durations for spells spanning multiple panel waves. To our knowledge the effects of DI and edit checks on reported amounts of receipt have not been examined, nor has the net effect on key estimates derived from combinations of survey items (Moore et al., 2009). Although the reduction in under-reporting of individual income sources with DI found in previous studies is substantial, it is not a priori clear to what extent this methodological improvement affects substantive conclusions about household income and related estimates. We contribute to this literature by examining to what extent DI and edit checks affect estimates of household income, estimated poverty rates and estimated rates of transitions into and out of poverty. For this purpose we use three waves of the British Household Panel Survey (BHPS), in which both edit checks and DI are used for the collection of non-labor income data, in a quasi-experimental way. The results suggest that traditional methods of interviewing under-estimate household income in the lower tail of the income distribution. The effects on estimated poverty status and transitions into and out of poverty are however small. In order to ascertain that the effects on survey estimates reflect an improvement of data quality, we further use an experimental study carried out in the context of the BHPS, which linked survey responses to administrative records, to examine to what extent DI reduces measurement error in the reporting of income receipt, in amounts of income, in the duration of receipt and in transitions onto and of income receipt.

5.1 Data

The BHPS and the experimental validation study

Our analyses are based on two data sources: the BHPS survey and an experimental validation study carried out in the context of the BHPS. The BHPS is a panel survey of the UK population that started in 1991 with a clustered and stratified address-based sample of 5,500 households. All household members aged 16+ are interviewed annually and followed as long as they remain in the UK. The individual response rates, conditional on response in the prior wave, are around 94% (RR1- AAPOR, 2008) for the waves included in our analyses (Taylor, Brice, Buck, & Prentice-Lane, 2009).

The experimental study was carried out using the former European Community Household Panel low-income sub-sample for Great Britain,

which had been surveyed as part of the BHPS since 1997. Funding for this sample expired in 2001 and the sample was interviewed once more in 2003 for methodological purposes. The methodological survey included a split-ballot experiment comparing independent and dependent interviewing for various sections of the questionnaire, including questions about non-labor income sources. In addition, respondents were asked for permission to obtain their records from the Department for Work and Pensions (the department in charge of administering cash transfers). The response rate for the methodological survey was 88.8% (RR1- AAPOR, 2008) of which 77.4% gave consent for linkage to the administrative records (Jäckle, Sala, Jenkins, & Lynn, 2004) and 74.1% of consenters were successfully matched to the records. Non-matched respondents were probably mainly respondents who had not received cash transfers during the time frame of interest, although some problems with the identifying information used for the linkage cannot be excluded (see Jenkins, 2008).

Based on findings from the experimental study, DI was implemented for several sections of the BHPS questionnaire as from 2006 (Jäckle, Laurie, & Uhrig, 2007). In addition, following recommendations by Lynn et al. (2012) edit checks were implemented as from 2005, to further reduce under-reporting. For comparability with the experimental survey, BHPS extension samples are excluded from the analyses.

Although the experimental validation and BHPS data are based on the same survey design, there are several differences between the surveys which are relevant to our analyses. The experimental validation data are based on interviews in 2001 and 2003. The 2001 survey used independent interviewing, while the 2003 survey experimentally allocated respondents to a DI treatment. In contrast, the BHPS used DI for all sample members, in both the 2006 and 2007 interviews. While both the experimental survey and the BHPS collected information about all components of non-labor income, validation data were only obtained for cash transfer data. In addition to using DI, the BHPS used edit checks for questions on cash transfer receipt in the 2005 (wave 15), 2006 (wave 16) and 2007 (wave 17) surveys, while the experimental study did not use any edit checks. There are furthermore differences in the length of the reference period between interviews (on average 17 months in the experimental data and 12 months in the BHPS data) and the sample composition (the experimental data over-represents low-income households).

Finally there are some differences in how the survey and the administrative records capture information about cash transfer receipt. First, a few transfer types included in the survey are not included in the records (Widowed Mother's Allowance, War Disability Pension, Council Tax Benefit). Second, in the record data some cash transfer types

(Disability Living Allowance, Child Benefit) are recorded as a single source, while the surveys collect separate information about different components (e.g. care component vs. mobility component). For comparability, the experimental survey data have been edited to reflect the data structure of the records.

The survey questions on non-labor income components

In the experimental survey, three versions of questions on non-labor income components were randomly assigned. With independent interviewing (INDI), respondents were shown a series of four showcards, listing a total of 34 potential income sources, and asked which of these they had received during the reference period: *"Please look at this card and tell me if, since September 1st 2001, you have received any of the types of income or payments shown, either just yourself or jointly?"* With proactive DI (PDI), respondents were reminded of each source they had reported in the previous interview and asked whether they had again received this: *"According to our records, when we last interviewed you, on <Date of interview>, you were receiving <Source>, either yourself or jointly. For which months since <Month of interview> have you received <Source>?"* Respondents were then asked the INDI version to catch any new income sources they had not previously reported. With reactive DI (RDI), respondents were first asked the INDI question. For any income sources reported in the previous but not the current interview, they were asked a follow-up question: *"Can I just check, according to our records you have in the past received <Source>. Have you received <Source> at any time since <Date of interview>?"* For each income source reported, respondents in all experimental conditions were then asked the same series of follow-up questions about the timing and amounts of receipt: *"And for which months since September 1st 2001 have you received ..... ?"*, *"How much was the last payment of ..... you received?"*, and *"What period did that cover?"* For the second and further income sources, it is possible for respondents to report that the amount was already included in the report for a previous income source.

The BHPS implemented edit checks as from 2005 for those cash transfers, for which answers to earlier questions in the same interview predict eligibility: Pension Credit, Disability Benefits, Income Support, Job Seekers' Allowance, Child Benefit and Housing Benefit. For these questions respondents who had not reported receipt, but whose responses to prior questions in the interview suggested that they might be eligible, were asked an edit check question. For example, respondents above the State retirement age who had not reported receipt of a State pension were asked *"Can I just check, do you currently receive the State*

Retirement Pension?" From 2006 onwards, the RDI version of the non-labor income questions tested in the experimental survey was implemented. (The BHPS questionnaires can be viewed at http://www.iser.essex.ac.uk/survey/bhps). Although the BHPS data are not experimental, the public release file identifies which income sources were reported in response to the initial INDI question, which in response to the in-interview edit checks, and which in response to the RDI follow-up questions, enabling a quasi-experimental comparison of the effects of the interviewing method on responses and estimates.

For analysis purposes, we group the income sources into four components of non-labor income: cash transfers, private pensions, other transfers and investments. This grouping corresponds to the derived income components provided with the BHPS public release file. Table 10 lists the sources corresponding to each component. Labor income also contributes to household income. As DI and edit checks were not used for labor income, we do not examine this component separately, although it is included in the measures of household income we derive. Table 11 documents the number of cash transfer, pension, other transfer and investment sources reported in waves 15-17, either in response to the initial INDI question, the edit checks, or the reactive DI follow-up question for the BHPS respondents. Based on the enumerated sources, the timing, amounts and nature of receipt, we computed three versions of household income based on only the INDI-reports, the INDI and edit checks reports and finally the INDI, edit checks and RDI-reports.

Table 10: Components of Non-Labor Income

Cash transfer Income	- 4 types of national insurance pensions and tax credits - 10 types of disability related cash transfers and tax credits - 2 types of income support - Housing Benefit - Council Tax Benefit - Job Seekers Allowance - Child Benefit - Maternity Allowance - Working Families Tax Credit - Child Tax Credit
Pension income	- 3 types of private pensions
Other transfers	- education grants - sickness insurance - maintenance/ foster allowance - payments from trade unions/friendly societies - payments from absent family members - other payments
Investment income	- rent from boarders/lodgers - rent from other properties

Table 11: Number of Income Sources Reported in the BHPS

		Cash transfers	Pensions	Other transfers	Investment
Wave 15	INDI	8088	1717	426	274
	Checks	117	–	–	–
Wave 16	INDI	8170	1776	515	323
	Checks	165	–	–	–
	RDI	615	121	55	39
Wave 17	INDI	7895	1846	501	302
	Checks	157	–	–	–
	RDI	506	94	49	423

Notes: Number of respondents at wave 15:8,538, wave 16:8,484, wave 17:8,322.
INDI: Independent Interviewing, checks: within-wave edit checks, RDI: Reactive Dependent Interviewing, based on all original BHPS sample members

Table 12: Sample Sizes in the Experimental Validation Data

	Respondents consented to linkage	Income sources in records	Income sources in survey
INDI	262	374	338
PDI	263	391	376
RDI	274	407	401

Notes: INDI: Independent Interviewing, RDI: Reactive Dependent Interviewing, PDI: Proactive Dependent Interviewing

Table 12 documents the number of respondents in each of the experimental treatment groups who consented to the record linkage, as well as the number of cash transfers reported in the 2003 survey and corresponding administrative records.

5.2 Results

Effects on estimates of household income and poverty

To examine whether DI and edit checks affect estimates of household income, we use waves 15 to 17 of the BHPS. Table 13 shows estimates of the equivalized annual household income distribution for the population of Great Britain. The estimates are based on all members of surveyed households, adjusted for differences in household size using the McClements equivalence scale (Taylor et al., 2009) and weighted for non-response. The first column indicates the estimated income percentiles including only income sources reported in response to the INDI questions. The second column indicates by how much the income percentile changes when income sources reported in response to the edit checks are included. For waves 16 and 17, the third column indicates by how much the INDI estimate changes if sources reported both in response to the edit checks and the RDI follow-up questions are included. Edit checks have a

considerable effect, increasing estimated income percentiles below median income, for example increasing the fifth percentile by 6% at wave 16. RDI has an additional effect, beyond that of the edit checks, increasing the income estimate by a further 4% points to 10%. The effects of RDI and edit checks are largest for people in the lowest percentile, monotonically fall across percentiles, and are zero, or close to zero for all percentiles above the median. As a result, the effects of the edit checks and RDI on median income are small: when the edit checks reports are included the median income increases by less than 0.3% in each of the three waves, and by a further 1% at waves 16 and 17, when responses to RDI are included.

Table 13: Estimated Distribution of Equivalized Annual Household Income

	Wave 15		Wave 16			Wave 17		
Percentie	INDI (£)	% change (INDI + Checks)	INDI (£)	% change (INDI + Checks)	% change (INDI + Checks + RDI)	INDI (£)	% change (INDI+ Checks)	% change (INDI + Checks + RDI)
1	1609	49	436	183	210	842	95	137
2	3740	6	2780	31	35	3205	26	32
5	6047	5	6073	6	10	6385	4	7
10	8353	1	8549	2	6	8630	2	6
25	13594	0	13881	0	3	13847	1	4
50	25192	0	25267	0	2	25594	0	1
75	40921	0	41106	0	1	42472	0	1
90	57812	0	58602	0	0	61828	0	0
95	71091	0	73977	0	0	75365	0	0
98	89872	0	93273	0	0	96199	0	0
99	107793	0	109815	1	1	115145	0	0
Median	25192	25250	25267	25324	25668	25594	25625	25933

Notes: Based on all enumerated household members, wave 15:11,700, wave 16:11,611, wave 17:11,374.
INDI: Income derived from sources reported in response to Independent Interviewing, INDI+Checks: Income derived from sources reported in response to both INDI and edit checks, INDI+Checks+DI: Income derived from sources reported in response to INDI+Checks and Reactive Dependent Interviewing.

The results suggest that DI and edit checks increase estimates of household income at the lower end of the income distribution, where cash transfers, pensions and other transfers are likely to represent a major component of total income. For households with higher levels of income, these sources are likely to be less important.

Effects on estimated poverty rates

To examine whether DI and edit checks affect estimated poverty rates, we use poverty classifications based on the BHPS for waves 15 to 17. A poverty threshold frequently used for official statistics is 60% of median current income (Department, 2008). We use the same poverty threshold, but use annual household income instead of current income, as we are interested in the net effects of the edit checks and RDI on the timing and amounts of receipt during the year. Böheim and Jenkins (2006) show that differences between poverty indicators based on current and annual incomes are very small.

Table 14: Estimated Poverty Rates

		% of individuals classified as 'poor'	% classified as 'poor' with INDI, but not with the additional checks	% classified as 'poor' with additional checks, but not with INDI
Wave 15	INDI	18.6	–	–
	INDI + Checks	18.5	0.8	0.0
Wave 16	INDI	18.9	–	–
	INDI + Checks	18.8	0.9	0.1
	INDI + Checks +RDI	18.4	4.2	0.4
Wave 17	INDI	18.4	–	–
	INDI + Checks	18.2	1.2	0.0
	INDI + Checks +RDI	17.9	3.9	0.3

Notes: Based on all enumerated household members, N at wave 15:11,700, wave 16:11,611, wave 17:11,374. Poverty threshold defined as 60% of median annual equivalized household income. Estimates are weighted to adjust for non-response.

The results in Table 14 suggest that the edit checks and DI reduce the estimated percentage of individuals living in households with annual incomes below the poverty threshold. The reduction in poverty estimates is small, but consistent across waves. It reflects the findings from Table 13, that the additional checks mainly affect estimated household incomes in the lowest percentiles, while the median income hardly shifts. Correspondingly, the second column indicates that most changes in poverty classification occurs for individuals that are classified as 'poor' with INDI, but are no longer 'poor' when sources reported in response to the additional checks are added to their income. This occurs for example for 4.2% of the respondents in wave 16, when the design with DI and edit checks is compared to the INDI design. Some households classified as 'not poor' with INDI, however become 'poor' with the additional checks and are shown in the third column, this occurs for example for 0.4% of respondents in wave 16. These are households whose income is only just

above the poverty threshold based on the INDI data and do not report any additional income sources in response to the checks or DI. Therefore, they slip just below the poverty threshold when this includes the edit check and DI responses.

Effects on estimated poverty transitions

To examine whether the additional checks affect the longitudinal consistency of poverty classifications across waves, we again use the BHPS data.

Table 15 shows the transitions for each pair of waves (15 to 16, 16 to 17), based on the INDI data only, adding the edit check data, and further adding the RDI data. Because we use a relative measure of poverty, changes in poverty status across waves could be caused both by the shift in the median income and a change in an individual household's income. For example, according to the INDI data, 76.3% of individuals were living in non-poor households in both waves 15 and 16, while 5.6% entered poverty during this period. The edit checks and DI lead to a small increase in the percentage of persistent non-poor people. Neither edit checks nor DI have much effect on transitions rates into- and out of poverty however. This result is surprising, since we would have expected DI at least to increase the consistency of responses across waves, and by implication to reduce changes in household income and resulting poverty status across waves.

Table 15: Estimated Transition Rates into and out of Poverty

	Transition type	INDI	INDI + edit checks	INDI + edit checks + RDI
Wave	Persistent non-poor	76.3	76.4	–
15-16	Persistent poor	13.1	13.0	–
	Transition into poverty	5.6	5.7	–
	Transition out of poverty	4.9	5.0	–
Wave	Persistent non-poor	76.3	76.5	77.0
16-17	Persistent poor	13.2	13.2	12.9
	Transition into poverty	5.0	4.8	4.8
	Transition out of poverty	5.5	5.5	5.3

Notes: Based on all enumerated household members, N wave 15-16:10278, wave 16-17:9692. Data are weighted to adjust for attrition.

To summarize, we find that both edit checks and DI increase estimated household income in the lower tail of the income distribution, but neither method has much effect on estimated poverty classifications or

transitions into and out of poverty. The next section examines whether the change in estimated household income is likely to reflect an improvement in data quality, by examining the effects on a series of indicators of measurement error which are related to household income.

Effects of DI and edit checks on measurement errors

To examine the effects of check questions on measurement error, we use the validation data. As outlined in Section 2.1, there are several differences between the experimental validation data and the BHPS data. To be able to draw any conclusions about how the results from the validation study relate to the estimates from the BHPS presented in the previous section, we have checked that the effects on reporting are similar in the two surveys. First, we checked that the effects of RDI on reporting cash transfer receipt are similar in both data sources. Second, we checked that the effects of edit checks, used only in the BHPS survey and only for cash transfer incomes, are in the same direction as the effects of RDI. Third, we checked that the effects of RDI on reporting private pensions, other transfers and investments in the BHPS are similar to the effects on reporting cash transfer income. See Appendix B for a full comparison of the validation and BHPS data.

The results indicate that the differences in reporting patterns between INDI and RDI are similar in the experimental survey and the BHPS. We therefore conclude that the results from the experimental data presented in the following section are likely to also hold in the BHPS data. In addition, the effects of the edit checks on reporting are in the same direction as the effects of RDI. We therefore further assume that the edit checks are likely to have similar effects on measurement error as RDI.

Effects of DI on measurement error in receipt of income sources

The first step in examining the effect of DI on measurement error focuses on individual reports of receipt of income sources. We compare responses to the experimental survey with individual register data. For each potential income source, we derive indicators of whether or not the source was received at any point during the reference period. Separate indicators are derived for the survey and the record data and used to classify all potential income sources for each respondent: *true negatives* are income sources which were neither received according to the survey, nor according to the records; *true positives* are income sources which were received both according to the survey and the records; *false negatives* are income sources which were received according to the records, but not reported in the survey; *false positives* are income sources

which were not received according to the records, but reported in the survey. To account for the possibility that respondents may report income sources which are recorded in the name of a different household member in the record data, income sources are counted as 'true positives' if there is a record for the source in the name of another household member.

Table 16 indicates the number of potential income sources which are classified as true/false positives/negatives. Assuming that the record data represent the true values, we interpret 'false negatives' as indicators of under-reporting, and 'false positives' as over-reporting. The last two columns indicate the corresponding error rates: the false negative rate is the number of false negatives as a proportion of sources received according to the records; the false positive rate is the number of false positives, as a proportion of the sources not received according to the records.

Table 16: Effect of DI on Measurement Error in Income Receipt Reported by Individuals

	Sample sizes (N)				Error rates (%)	
	True negative	False negative	False positive	True Positive	False negative rate	False positive rate
INDI	3257	73	26	312	19.0	0.8
PDI	3246	60	25	351	14.6	0.8
RDI	3377	58	30	371	13.5	0.9

Notes: Sample sizes are based on the number of respondents documented in Table 11, multiplied by 14 potential income sources. Columns are defined in the text. False positives are counted as true positives, if the source is recorded for a different household member in the record data INDI: Independent Interviewing, RDI: Reactive Dependent Interviewing, PDI: Proactive Dependent Interviewing.

As in previous studies focusing on the effect of DI on reported receipt (e.g. Lynn et al., 2012), the results indicate that the main type of error is under-reporting: with INDI 19.0% of sources recorded in the records are not reported in the survey. Over-reporting hardly occurs: with INDI less than 1% of sources not received according to the records are reported in the survey. DI reduces, but does not eliminate under-reporting: the false-negative rate falls to 14.6% with PDI and 13.5% with RDI. DI does not have any effect on over-reporting: the false positive rates are similar across treatment groups. We therefore conclude that the increase in the reporting of income sources with DI represents a reduction in net measurement error of receipt of income sources.

Effects of DI on measurement error in the amounts of non-labor income

To test whether DI affects measurement error in the amount of income, we again compare the survey reports to the individual records. For each source we derive the amount of the last payment during the reference period according to the survey and according to the records. The amounts are standardized to weekly amounts, for comparability with the format in which they are recorded in the administrative data. We then calculate the error in amounts of receipt as the difference between the survey and the record amount. In the final step, we calculate the mean error over all cash transfers and respondents. The analysis excludes information about Housing Benefit. The record data for this income type stem from a different source than for all other cash transfer types. While the data on dates of receipt from this source appear consistent with the survey data, the data on amounts of receipt contain large amounts of inconsistencies which we have not been able to resolve.

Table 17: Effect of DI on Measurement Error in Amounts of Receipt

	Mean difference in amounts between survey and records	95% C.I.	
INDI	-4.6	-9.12	-0.15
PDI	-2.6	-6.16	0.90
RDI	-5.9	-9.86	-1.92

Notes: The base includes all sources reported in either the survey or the register or both: INDI sources N:278, PDI sources N:290, RDI sources N:306.

The results in Table 17 indicate that with INDI weekly non-labor income is under-reported by on average £4.6. With PDI the mean error is reduced slightly to £2.6, but with RDI it unexpectedly increases to £5.9. This suggests that although DI reduces under-reporting of receipt, it does not help respondents to improve their reporting of the amounts received.

Effects of DI on measurement error in duration of receipt

To assess the effects of DI on measurement error in reported duration of receipt, we again compare the survey and administrative data. For each income source we calculate the error as the difference between the number of months of receipt according to the survey and the records. We then calculate the mean error over all income sources and respondents.

Table 18: Effect of DI on Measurement Error in Months of Receipt

	Mean error in months of receipt	95% C.I.	
INDI	-0.92	-1.44	-0.39
PDI	-0.32	-0.74	0.10
RDI	0.01	-0.40	0.43

Notes: The analysis is restricted to receipt between Sept. 1st 2001 and Sept. 1st 2002, for comparability with the BHPS data. The base includes all sources either reported in the survey, or recorded in the administrative data, or both, but excludes true negatives. INDI sources N: 392, PDI sources N: 423, RDI sources N: 440.

The results in table 18 suggest that with INDI receipt is under-reported by on average 1 month. With PDI and RDI the mean error is closer to zero and, according to the 95% Confidence Intervals, no longer significantly different from zero. This suggests that DI reduces measurement errors in reported duration of receipt of cash transfers.

Effects of DI on measurement error in transitions of cash transfer receipt across waves

Finally we evaluate whether DI reduces measurement error in reported transitions of cash transfer receipt across waves. We classify each potential income source for each respondent according to the type of transition between the 2001 survey and the 2003 survey as continued non-receipt, continued receipt, transition off receipt, and transition onto receipt. Each potential income source is classified separately based on the survey data and based on the record data. We then compare the transition types derived from the survey and records to identify errors in transition classifications.

Overall, with INDI the transition type is misclassified for 4.0% of potential income sources. With PDI the error rate is 3.4% and 3.9% with RDI, suggesting that DI does somewhat reduce overall error rates in transitions.

Since DI was only used in the 2003 interview, the interviewing method cannot have affected the wave 2001 status. Therefore Table 19 focuses on errors in the classification of transition types, conditional on the 2001 status being reported correctly in the survey. The first column indicates the transition type according to the records. The second column indicates the number of potential income sources to which this transition type applies. The third column indicates the percentage of income sources for which the 2003 status was misclassified, resulting in an error in the transition type.

Table 19: Effect of DI on Measurement Error in Transitions onto and of Cash Transfer Receipt, Conditional on Correct Classification in the 2001 Survey

	INDI		PDI		RDI	
Transition in Records	N	% misclassified	N	% misclassified	N	% misclassified
Persistent non-receipt	3230	0.5	3231	0.3	3348	0.5
Transition on	49	20.4	50	36.0	57	38.6
Persistent receipt	284	11.3	305	2.0	317	3.2
Transition off	16	0.0	5	0.0	15	13.3

DI does not have any effect on continued non-receipt: the error rates are close to 0.5% for all treatment groups. Also as expected, error rates for continued receipt fall from 11.3% with INDI, to 2.0% with PDI and 3.2% with RDI. This implies a reduction in the over-reporting of transitions off income receipt.

For transitions onto cash transfer receipt, however, DI unexpectedly increases the error rate, from 20.4% with INDI to 36.0% with PDI and 38.6% with RDI. Transitions onto receipt are more likely to be misclassified as continued non-receipt with both DI methods. As non-receipts in the previous interview do not trigger any DI-questions, the increased error rate in the DI conditions is a surprising finding.

Transitions off cash transfer receipt tend to be reported correctly in the INDI data (but the number of transitions is very small). RDI increases the likelihood that transitions off receipt are misclassified as continued receipt. This could be due to respondents falsely confirming a receipt status presented to them from the previous interview. If this were the case, we would however also expect this type of error to increase with PDI, which it does not.

A potential cause of the findings for transitions onto and off receipt could be found with the interviewers. In both DI-designs, they might be more focused on reducing errors in continued receipt than in picking up transitions onto and off receipt (Sala, Uhrig, & Lynn, 2009). Since the number of transitions onto and off receipt is small, we would however interpret these results with caution.

We conclude that DI affects measurement error in income receipt transitions between two consecutive waves. DI reduces under-reporting of continued receipt. A secondary effect is that DI potentially increases the spurious consistency in receipt status across waves, both by

increasing the misclassification of transitions onto receipt as continued non-receipt, and of transitions off receipt as continued receipt.

5.3 Conclusion and discussion

The motivation for this study was to examine what effect methodological innovations that are expected to reduce measurement error have on substantive survey statistics. In this case, it matters in practice whether or not income data elicited by edit checks and DI- questions matters are used to estimate household income. Methodological studies designed to evaluate the effects of alternative data collection methods on data quality often only examine the effects on answers to individual survey questions. We believe that an evaluation of the impact on data quality needs to relate to the actual uses of the survey data. In this spirit, we examine the effects of edit checks and DI on derived estimates, and subsequently whether these effects reflect a decrease in measurement error. For this purpose we exploit a unique combination of data sets: we use data from the BHPS, a large-scale panel survey which has implemented edit checks and DI for questions on non-labor income components in a quasi-experimental way, and from an experimental validation study based on the BHPS survey design. We use the experimental study to assess the effects of DI on different aspects of measurement error, and the BHPS data to assess the effects of DI and edit checks on estimates of household income and poverty.

The results suggest that both the edit checks and DI increase estimates of total household income in the lower tail of the income distribution. Neither method has much effect on estimated poverty rates or estimated rates of transitions into and out of poverty. The increase in household income reflects an increase in data quality: DI reduces under-reporting without affecting over-reporting; DI reduces under-reporting of months of receipt, erroneous transitions of income receipt and under-reporting of continued receipt across waves. The effects of edit checks are in the same direction and we therefore assume that they also reduce measurement error. The effects of DI on measurement error are in our view considerable. For example, the under-reporting rate is reduced by about 29% with reactive DI compared to independent interviewing. The effects on estimates of household income and poverty are arguably small. This suggests that while edit checks and DI may have large effects on measurement error in responses to individual survey questions, the combined effects, in this case over different survey items and different household members, may be small. This conclusion is however open to interpretation, since a reduction in the estimated poverty rate by a mere

0.5 percentage points affects around 300,000 individuals in the population of Great Britain. Moreover, DI and edit checks were only used for the non-labor components of income. Measurement errors in labor income, which contributes greatly to household income, are not accounted for.

There are a number of issues, regarding both the effects of DI and edit checks, and the mechanisms through which these methods work, which in our view warrant further attention. Reactive and proactive DI have rarely been compared experimentally. Although the results presented in Table 16 to Table 19 suggest some differences between the two in their effects on the various aspects of measurement error, they are small. We do however find evidence for the fact that both DI strategies are in particular successful to reduce measurement errors due to underreporting of continued receipt in two waves. This comes at the price of increased errors in situations where respondents experience a transition onto receipt.

The reason why reactive DI was implemented in the BHPS was that this made it possible to maintain comparability with the previous 15 waves of data collection, in which independent interviewing was used. The responses given to the independent question can still be identified and, for comparisons with previous waves, the responses to the reactive follow-up can simply be ignored.

Our ability to compare the effects of edit checks and DI were limited by the fact that both were always used in combination, with the edit checks always preceding the DI checks. Their relative effects may be quite different if compared individually or in different order. Since the edit checks do not require feeding forward information from previous interviews, they can be used in cross-sectional surveys and are cheaper to implement than DI. Their use is however restricted to income sources for which there are questions earlier in the questionnaire which are very good predictors of eligibility.

We assume that the changes in estimates due to edit checks and DI represent reductions in bias due to measurement error. The effects on total survey error may however be more complex. The official poverty rates for the UK for 2005, 2006 and 2007 are 21.7%, 22.2% and 22.5% (Department, 2008). These rates are not strictly comparable to our estimates, because our analyses are restricted to Great Britain and based on annual household income before taxes, while the official rates are based on current disposable income. Both are however based on equivalized household income, using 60% of median income as the poverty threshold. Although the rates themselves are not comparable,

there is nonetheless a key difference: the official statistics suggest an increasing trend over the three years, while our estimates including DI suggest a slightly decreasing trend. One explanation for this effect could be that measurement error and attrition error may be counterbalancing each other, and that when one is reduced, the effect on total survey error worsens. In the BHPS individuals with incomes in the bottom 40% of the distribution are more likely to drop out of the panel (Lynn, Buck, Burton, Laurie, & Uhrig, 2006). The under-representation of the low-income population may to some extent be counter-balanced by under-reporting of income sources. When measurement error in household income for these groups is reduced, the effect on total survey error may therefore be worse. This however remains a conjecture, since evaluating the relative contributions of measurement and attrition error would require information about non-respondents.

The long-term effects of DI on data quality have not been assessed. The ability of DI to reduce under-reporting is limited by the fact that the respondent can only be reminded of income sources reported in the past. Over time, as under-reporting is reduced with the help of DI, more information becomes available about which the respondent can be prompted. As a result the reduction of measurement error may well increase, since respondents can be reminded of a larger proportion of the sources they have received in the past.

The extent of measurement error in independent survey questions is presumably affected by the question format. The shortcut method of using show cards instead of separate yes/no questions about the receipt of all potential income sources presumably leads to more under-reporting. On the other hand, the shorter interview time reduces respondent burden, which could lead to less measurement error using the show cards. This trade-off between cost savings in terms of questionnaire time and measurement error has to our knowledge not been assessed.

Finally, we have not touched on the question through which mechanisms DI and edit checks work, i.e. which types of sources are most likely to be misreported, by which types of respondents, and how the edit checks work for these different groups (see Lynn et al., 2012). We have also not touched on the question how these methods could further be improved. These should focus further on the reduction of underreports (with DI), but also capture new receipts. This can be done by extending the use of edit checks by incorporating more factual questions that predict eligibility for income receipt. Measurement error in household income was reduced by our study design, but there is room for further reductions in error with more impact on substantive conclusions.

6 Chapter 6: Change? What change?

An exploration of the use of mixed methods research to understand longitudinal measurement variance[10]

The strength of panel survey studies in the social sciences is their ability to measure change over time at both the microlevel and the aggregate level. To measure change reliably, researchers use the same questionnaires for every measurement occasion or 'wave'. Apart from boring the respondents, repeated questionnaires have a serious methodological downside.

Consider a longitudinal study about gender perceptions among secondary school students. As puberty progresses, adolescents generally not only become more interested in the other gender, but the nature of their interest changes as well (Galambos, Almeida, & Petersen, 1990). If the same questionnaire is used to measure gender attitudes during adolescence, certain questions lose their importance and meaning with time, while others become more relevant to the respondents. Golembiewski, Billingsley and Yeager (1976) introduced this problem as 'gamma change': respondents redefined or reconceptualized the phenomenon of interest in the time between two measurements (Terborg, Howard, & Maxwell, 1980). Gamma change makes the comparison of levels of gender attitudes throughout puberty problematic (Saris & Gallhofer, 2007). In a qualitative research context, Morse and Niehaus (2009) refer to gamma change as the 'butterfly phenomenon'.

This paper deals with the problem of what change means when the concept of interest itself changes over time. Statistical methods have been developed to evaluate the stability of the meaning of a social construct. These methods centre on the concept of measurement invariance, which assumes that the relative importance of the indicators measuring the social construct is stable over time. Measurement invariance is tested with confirmatory factor analyses (CFAs), in which factor loadings and intercepts(and ideally also variances) are tested for equality across different samples (Vandenberg, 2002). To examine whether questionnaires can be used across groups, measurement invariance is tested across genders or nationalities, for example (Schmitt & Kuljanin, 2008).

Researchers often find large differences in factor structure, loadings, or intercepts, which make it impossible to compare levels for the factor scores across groups. In a longitudinal context, we can test measurement

10 This chapter was co-written with Hennie Boeije and Gerty Lensvelt-Mulders and is forthcoming in Methodology.

invariance to assess an instrument's reliability and validity over time. For instance, Vandenberg and Self (1993) examined orientation to paid employment among new employees on their first day of work and at 3 months and 6 months after they started working. Predictably, the employees' orientation towards their employer was far more diffuse on the first day than at later times. This resulted in lower correlations between questionnaire items and a diffuse factor structure in wave 1 and much higher factor loadings in the later waves. Furthermore, Schmitt and Kuljanin (2008) reviewed 75 recent papers that tested measurement invariance and found that full measurement invariance seldom holds. These findings suggest that the meaning that respondents perceive in survey questions often changes over time or between subgroups of the study population (Schmitt & Kuljanin, 2008; Vandenberg & Lance, 2000).

Solutions for research in which measurement invariance does not hold are not readily available. One option is to change the questionnaire and adjust, remove, or add questions. This option is unattractive because it would damage the comparability of the measurements in the different waves. Futhermore, adapting the study design to add new questions during the study is infeasible because measurement invariance is always assessed post hoc. The option we prefer is a combination of quantitative and qualitative methods, referred to as mixed methods.

Morse and Niehaus (2009 p. 14) define a mixed-method design as 'a scientifically rigorous *research project*, driven by the inductive or deductive *theoretical drive*, and comprised of a qualitative or quantitative *core component* with qualitative or quantitative *supplementary component(s)'.* This definition stresses that mixed-method research involves a primary method and a supplemental component that is partially complete and that is not conducted rigorously enough to stand alone or to be published by itself (Morse & Niehaus, 2009).

For our research project, we chose a mixed-method design in which we conducted the core and supplemental projects simultaneously. We added qualitative methods (i.e. qualitative interviews) to the quantitative survey so the results of qualitative methods provide a context that enables interpretation of the results of the analysis of measurement invariance. Qualitative methods are thought to tap different dimensions of social realities than quantitative methods (Creswell & Plano Clark, 2007; Mason, 2006). In doing so, they provide a valuable background for understanding the findings from analysis of measurement invariance. While quantitative instruments are never changed during the study, qualitative methods are flexible and can be adjusted to the field. Indeed, one of the principles of qualitative research methods is to tune in to the field and to change over

time (Jorgensen, 1989). In short, the use of mixed methods allows us to observe change from different angles.

This paper reports our testing of longitudinal measurement invariance with data from two quantitative surveys. We have acquired qualitative insights into structural changes of the variables from several supplemental qualitative interviews. As an example, we study the development of the construct 'study motivation' among first-year psychology students.

Few panel studies have incorporated a mixed-methods component into their study, and the use of interviews to deepen insights from analysis of measurement invariance is novel. Knowledge about the nature of measurement variance and change will add to our knowledge on the quality of estimates of change in panel studies. Increasing data quality in panel studies is important because the resulting data are frequently used to inform policy interventions. Our study provides insight into the nature of respondent characteristics that are amenable to longitudinal measurement with both quantitative and qualitative methods.

6.2 Research Methods

Overall Design and Sample

Our study started at the beginning of the academic year in September 2007. During their first week at Utrecht University, we invited 444 freshman psychology students to participate in the first survey of the panel study. We e-mailed them an individualized URL that led them to a website where they could join the first survey, presented in the form of computer assisted self interviewing (CASI). Our goal was to identify determinants of study success; hence the study focused on collecting demographic, psychological, and social characteristic information.

Of the 298 respondents who completed the first survey (hereafter referred to as wave 1), 69% responded initially (RR 1 - AAPOR, 2008). This response rate was relatively low partly because we e-mailed students at their new university e-addresses, which, we later discovered, some students were not yet using. The response rate for wave 2 in March and April 2008, conditional on the previous wave, was 80%[11] (RR1 - AAPOR, 2008). 81 per cent of the students was female, and the mean age for the students was 19.8 years when starting university. Three quarters of respondents finished high school the year before entering university, and

[11] Due to the longitudinal nature of the study and mortality, the response rates (RR2) are difficult to compute exactly.

their average grade over the entire first year of studies on a 1 (lowest) to 10 (highest) scale is 6.98. All these sample statistics are common for Dutch first-year psychology students.

Of the wave 1 participants who gave consent for a follow-up contact, we contacted ten by phone or e-mail and asked them to participate in a qualitative follow-up study. Some never replied, while others did not want to be interviewed after all, did not show up, were late, or were in a hurry. Finally, we interviewed four wave 1 participants. We analysed the data and the qualitative interviews separately and discussed the results. After wave 2 (April 2008), we invited ten students for a qualitative interview, and again, four participated. We compared the findings from waves 1 and 2 for similarities and differences. The sample for the qualitative interviews is described in Table 20. The interview participants are individually referred to as R1 to R4 (wave 1) and R5 to R8 (wave 2).

Table 20: Participants in qualitative interviews

		WAVE 1	
	Gender	Age at wave 1	Average grade in first year
R1	Female	18	7.3
R2	Female	18	7.0
R3	Female	19	7.5
R4	Male	19	7.1
		WAVE 2	
R5	Female	19	7.1
R6	Female	20	6.8
R7	Male	32	7.6
R8	Male	21	6.3

Quantitative measures

Study motivation is deemed to be one of the most important determinants of study success (Hidi & Harackiewicz, 2000). Our study motivation questionnaire was adapted to first-year students on the basis of Ryan and Connell's (1989) work. It included 18 statements for the students to rate as 'not true at all', 'not true', 'sort of true', 'true' or 'very true' for themselves. The statements reflected external reasons for studying ('That's what I'm supposed to do' or 'I don't want others to get mad at me') and internal reasons ('I want to learn new things' or 'I enjoy it'). Appendix C shows the full survey questions and answer categories.

For none of the items were the distributions very skewed or kurtose (< 2.0) and the median and mean is about 3 for most of the variables. Missing data on any of these individual items were dealt with using the Full Information Maximum Likelihood (FIML) approach, as implemented in AMOS 16.0 software (Arbuckle, 2007). 37 students did not complete the study motivation scale in both wave 1 and wave 2, and were deleted from the analysis.

We also used AMOS 16.0 to process data from waves 1 and 2 in a CFA to determine whether measurement invariance held across waves 1 and 2. Testing measurement invariance involves testing various hierarchical models while increasingly constraining model parameters. Figure 3 shows the two-factor structure that illustrates the steps for establishing longitudinal measurement invariance for study motivations.

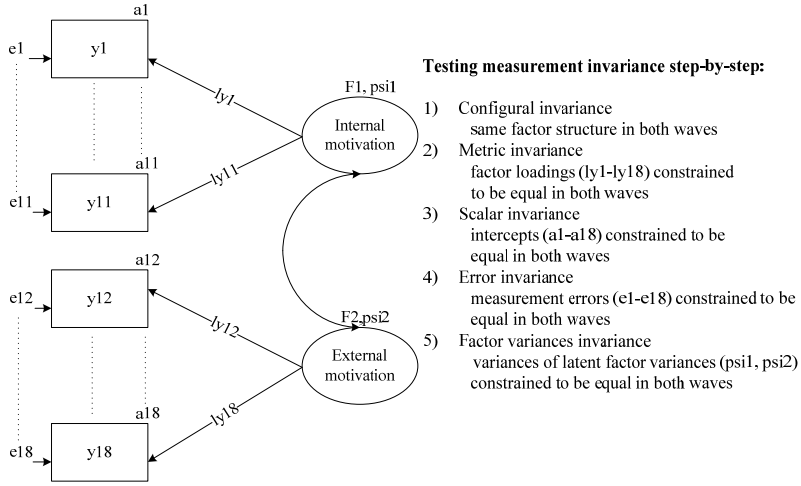

Figure 3: Testing measurement invariance in a two-factor structure for study motivation

Step 1 in testing measurement invariance consists of establishing whether the same hypothesized factor structure holds in both waves. The factor structure (Figure 3) is imposed on the data from both waves to see whether the same items load on the same factors. If this model were to fit, the values of the factor loadings (ly1 – ly18) are constrained to be equal in both waves (step 2). This will show whether the meaning of two latent constructs can be compared over time. If the constrained model were not to fit significantly worse than the unconstrained model, the intercepts (a1 – a18) are next constrained to be equal in both samples. The presence of both equal factor loadings and equal intercepts

establishes *scalar invariance* (step 3), which holds that both the meanings and levels of the latent factors can be meaningfully compared across time. In steps 4 and 5, the measurement error variances and, finally the latent factor variances are then constrained to be equal[12].

By testing measurement invariance, we answer three questions that relate to changing notions of the concept of interest. First, we determine whether students have a clear motivation to study when they start studying and halfway through their first year (factor structure). Second, we determine whether, in the interim, there was both individual and aggregate change in the relative importance of the different survey questions for measuring study motivation (factor loadings). Third, we assess any levels of change for any of the indicators (intercepts).

Qualitative measures

We conducted semistructured interviews with a topic guide. The main topics in wave 1 were the motivation to study psychology and respondents' study expectations. In wave 2, the focus was still on their motivation to study psychology, their experiences during the past year, the feedback on their performance, and future choices about elective courses in their psychology studies. We derived the topics from the literature and our educational experience with students, and we drew inspiration from the items in the survey questionnaire. The interviews lasted between 15 and 40 minutes. They were recorded and fully transcribed.

We used Maxqda2007, software for qualitative data analysis, to carry out our analyses (Kuckartz, 2007). First, the wave 1 data were disassembled into fragments and coded with labels reflecting their meanings (Boeije, 2010). We continued the coding process to build a code system that consisted of main codes and subcodes (Appendix C). Next, we added and analysed the wave 2 interviews. This resulted in new codes and renewal of the code system. The numbers behind the codes in Appendix C indicate how often each code was assigned to the interview fragments. Meanwhile, we interpreted the data and wrote memos about upcoming ideas, preliminary findings, and conceptual development. Then, we compared the results of waves 1 and 2, focusing on the development in the interim. We modelled the data so that it became clear how the students' study motivation developed during their first year.

[12]Additionally, constraining factor means would show whether the latent means differ significantly across time as a sixth step. However, this is outside the scope of the current study.

6.3 Results

Quantitative Results

Motivation. Earlier studies of study motivation (Ryan & Connell, 1989; Vallerand & Blisonnette, 1992) have found a clear two-factor structure discriminating between external and internal study motivations. Although more complicated structures have been suggested, the distinction between extrinsic and intrinsic study motivations is widely recognized, while psychometric properties of the scale are good (for an overview, see Ryan & Deci, 2000).

To test whether this structure was present in our data, we carried out a two-factor CFA for the data from waves 1 and 2. This first step in establishing measurement invariance is called configural invariance (Figure 3). We tested whether the same factor structure appeared in both waves (Vandenberg & Lance, 2000). Table 21 shows that the free models for wave 1 ($X^2(134) = 365$, root mean square error of approximation (RMSEA) = .078, comparative fit index (CFI) = .77), wave 2 ($X^2(134) = 400$, RMSEA = .045, CFI = .73), and the initial combined model ($X^2(592) = 1531$, RMSEA = .041, CFI = .62) do not fit the data well[13]. However, the RMSEA for the combined model is acceptable (RMSEA < .05) according to rules of thumb for model fit (Kline, 2005).

[13] In a review of measurement invariance testing, Vandenberg and Lance, (2000) suggest that, in comparisons of invariance models across groups, absolute RMSEA values below .06 reflect excellent fit and stepwise changes in the CFI of more than −.02 represent a definite loss of fit.

Table 21: Results of measurement invariance tests for study motivation in waves 1 and 2

Study motivation	Invariance test	χ^2	df	CFI	RMSEA	$\Delta\chi^2$	Δdf	Model accepted?
1a. Free model, wave 1		386	134	.77	.078	-	-	No
1b. Free model, wave 2		400	134	.73	.045	-	-	Yes
1c. Combined waves 1 and 2	Configural invariance	1531	592	.62	.041	-	-	Yes
2. + Factor loadings constrained	Metric invariance	1586	608	.61	.042	55	16	No
3. + Means constrained	Scalar invariance	1648	626	.59	.042	61	18	No
4. + Error variances constrained	Error invariance	1678	644	.58	.042	30	18	No
5. + Factor variances constrained	Factor invariance	1709	644	.57	.042	31	2	No

Notes: the first six columns show the model fit statistics. The last three columns show the improvement of the model relative to the previously accepted model. CFI = Comparative fit index; RMSEA = root mean square error of approximation. Adding covariances between some of the error terms within the same study motivation factor would greatly improve the fit of all models for study motivation. We decided not to add any error covariances in order not to obscure the interpretation of the factor loadings. *n*(wave 1) = 261, *n*(wave 2) = 196

Step 2 in invariance testing is establishing metric invariance. If metric invariance is established, the factor loadings of the items of study motivation are equal over time. This implies that the concept of motivation can be compared over time. Because the model with equal factor loadings in the two waves is equivalent to the model with configural invariance, except for the constraints on the factor loadings, we can use the difference in fit ($\Delta\chi^2$) to determine whether the constrained model fits significantly better than the free model. A nonsignificant $\Delta\chi^2$ means that the constrained (more parsimonious) model does not fit significantly worse than the free model. In such a case, the parsimonious model is preferable to the less parsimonious one. Our results showed that the model with equal factor loadings fit the data significantly worse ($\Delta\chi^2(16) = 55$, $p \leq .01$, ΔCFI = -.02).

As the study motivation scale has already been extensively used, we believe that the lack of invariance is a sign of a problem with our sample. Students have unclear and changing notions of their motivation to study. Moreover, their motives cannot be clearly separated into external and internal study motivations. Steenkamp and Baumgartner (1998) argue that if full metric invariance does not hold, releasing the constraints on one or several factor loadings can establish partial invariance. In this paper, we are interested in the sources of measurement variance in study motivation between the two waves, so we did not continue to study partial measurement invariance. We however inspected the factor loadings of the study motivation scale from model 1c in

Table 21 (configural invariance) more closely to identify the sources of misfit.

Table 22 shows the factor loadings for study motivation from the factor model with configural invariance (same factor structure). This table generally reflects the higher factor loadings for study motivation in wave 2, which means that students showed a clearer understanding of their motivation to study. Both external and internal motivational components appeared among the items that performed better in the later wave: item6 ('I want my teachers to think I'm a good student'), item12 ('I want to understand the subjects') and item18 ('I enjoy it'). A few items for external motivation had lower factor loadings in wave 2; item7 ('I feel bad if I don't') and item8 ('I feel ashamed if I don't').

Table 22: Unstandardized and standardized factor loadings for the study motivation questionnaire for waves 1 and 2 from model 1c (configural invariance)

	Wave 1			Wave 2		
Item	Unstandardized factor loading	Standard error	Standardized factor loading	Unstandardized factor loading	Standard error	Standardized factor loading
External motivation						
1)	1.000	–	.22	1.000	–	.49
2)	2.789	.539	.65	1.143	.168	.70
3)	1.460	.351	.41	.472	.098	.43
4)	2.026	.462	.45	.764	.145	.48
5)	2.458	.482	.63	.743	.131	.63
6)	.418	.046	.11	.630	.059	.40
7)	2.533	.508	.59	.546	.121	.39
8)	2.809	.538	.67	.869	.156	.51
9)	2.320	.470	.58	.816	.146	.52
10)	2.027	.443	.49	.635	.138	.40
11)	1.882	.391	.54	.678	.129	.48
Internal motivation						
12)	1.000	-	.56	1.000	-	.78
13)	.915	.139	.59	.666	.068	.68
14)	.104	.249	.03	.50	.13	.29
15)	.720	.159	.36	.445	.079	.41
16)	1.095	.165	.60	.662	.067	.68
17)	1.804	.229	.81	1.093	.085	.85
18)	1.620	.209	.78	1.160	.089	.87

Notes: n(wave 1) = 261, n(wave2)= 196
All factor loadings are significant at $p < .01$

In order to further explore the factor structure, we finally used Exploratory Factor Analysis (EFA) to check for other possible sources of misfit. These exploratory analyses suggested a third and fourth factor

could be added in both wave 1 and 2. None of the items loaded highly on these additional factors however, further indicating that the factor structure of study motivation is diffuse. Some items do not load at all on the latent factors of study motivation, while other items only load highly in wave 2.

Because the factor loadings of study motivation in waves 1 and 2 were different, we have to conclude that the notion of study motivation is different in the two waves. From inspection of the factor loadings, we conclude that some intrinsic motives became more important in the concept of study motivation, while components of external motivation, more specifically feelings of introjection, became less important (Vallerand & Blisonnette, 1992).

Qualitative Results

A transition period. The qualitative data showed that the step from secondary education to university was large for these adolescents. Initially, they were overwhelmed with new impressions, new friends, a new town, and, first and foremost, a new learning system. They were uncertain about their choices to go to university and study psychology. After the first months, they reflected upon the subjects they had to learn and found out what they really liked. Above all, they had to choose a specialism within psychology: an important decision that would frame their future.

The biggest development from the beginning to the end of the first year was that the initial uncertainty was replaced by decisiveness to continue their study. The entire first year seemed to be used to determine whether they made the right decision. Students answered two questions for themselves: Do I like it? Can I do it?

Motivation. Studying psychology for one student was something she had always wanted, for others it was a choice between various studies. Only one student remarked that he just wanted to go to university and was not particularly interested in psychology.

Most of the students observed a difference between secondary education and university in concern for motivation and encouragement. They found out that they were not personally encouraged to study and that they had to start studying by themselves:

> 'You are on your own here, and you do it for nobody but yourself. […] So that is difficult; if you do not feel like studying, then you do not. I am not really motivated from within the study, you have to motivate yourself, and that is difficult for me.' (R1, wave 1, ♀)

During the wave 1 interviews, questions about the relevance of studying psychology and about their professional future popped up every now and then, but the students were busy with the here and now of their first weeks at university:

> 'I just do not know yet what I want to do with psychology. I am looking around a bit – what do I like, what is interesting. What do I think of this chapter and which specialization track fits this chapter best? [...] What do I find interesting to read about, what seems fun to do later on in a job? [...] I am busy with "now" and not with "later on". I am postponing it a bit, it will come later on.' (R2, wave 1, ♀)

During wave 2 at the end of the first year, the students felt reassured about their choice to study psychology. Three students acknowledged that they were hesitant when they started. It was reassuring to them that they found their study pleasurable and interesting:

> 'I hope the study will focus more and more and that I will be busy with the subjects that interest me most. My interest, that is the attractive part for me. I am really looking forward to the specialization track. I am really looking forward to focusing, and specializing in my own aims.' (R7, wave 2, ♂)

At the end of their first year, students must choose a specialism in a specific psychological direction, such as organizational psychology or clinical psychology. They weighed what they found interesting and how they would see themselves after finishing their study:

> 'I chose social psychology. When I started my study, I was thinking more about clinical psychology, since that was the reason that I chose psychology in the first place. Not because I considered myself a psychologist – or maybe I did. Right from the start I liked the social psychology chapters. I had doubts for a long time and wanted to combine the two. But then I realized that to become a clinical psychologist you do need certain subjects in your major, otherwise you will not get a job. And then I decided to go for social psychology.' (R6, wave 2, ♀)

In short, the qualitative findings showed that the students had realized their expectations about university and that studying psychology was the right choice. They chose a direction and after having made this serious choice, the participants were sure what they wanted. They were motivated to take courses and take part in activities geared towards their objectives. Motivation for students then means that they are willing to spend time studying because they have an interest in what they have to learn and consider their efforts relevant for their professional future.

Integrating qualitative and quantitative results

We now turn to an integration of the results from the closed-ended survey questions with the qualitative interviews. Our foremost finding in the survey data analysis of measurement invariance was that our measurements varied over time; the concept of 'study motivation' did change over time during the study.

The qualitative research showed that students had unclear motivations at the start of their psychology studies, which might explain the poor fit of the factor pattern to the data and the low factor loadings. This was to be expected; the students had no clear idea what to expect in their first weeks at university, let alone how to be successful students. Whether the study was 'fun' or 'interesting' was impossible for them to tell yet because they had not really experienced what being a student is like. From the qualitative interviews, we can see that the first 6 months provided the students with a reality check. Therefore, it was not surprising to find that the fit of the factor structure and the values of the factor loadings for study motivation were better in wave 2, albeit still not good. Table 23 summarizes the similarities and differences between the quantitative and qualitative findings.

A closer look at the factor structures for the concept 'study motivation' revealed that specific survey questions had greater factor loadings in wave 2 than in wave 1. Both external study motivations (wanting to be seen as a good student) and internal ones (it is fun and interesting) became more important indicators of study motivation. Some other items in turn became less important indicators ('I will feel bad or ashamed if I don't'). These differences represent a shift in the student's motivations to study. The reasons for studying during the first year can be distinguished as internal and external motivations. The importance of the indicators changed within both these domains.

In the qualitative study, we found that students went through a period of transition. This contextual information clarifies why some items became better or worse indicators for study motivation throughout the first year. Students reported that studying raised their interest in psychology. Apart from these similarities between the quantitative and the qualitative findings, the qualitative findings showed that the students were occupied with possible specializations and their professional careers.

Table 23: Summary of findings from the quantitative data (survey) and the qualitative data (interviews)

Survey findings	Interview findings
Start of studies (wave 1)	
Students had a multitude of motivations to study, which could not be captured clearly in two dimensions Factor loadings for study motivation were low; the factor model did not fit the data well Items that asked students about their opinion about studying psychology performed relatively poorly	Everything was new, and students were uncertain about their choice to study psychology and to go to university Here and now is most important; feelings about the relevance of studying psychology and expectations about their professional future lure in the background
End of first year (wave 2)	
The concept of study motivation had changed over the year. The factor pattern had become clearer, while factor loadings had generally become higher Specific internal motivations (fun and interesting) and external motivations (because I want to be seen as a good student) had become more important as indicators of study motivation and factor loadings than in wave 1	Students had a clear image of what studying psychology means and what has their interest They had a clearer image of their future plans (in their studying and professionally) Students took part in activities geared towards their study objectives

6.4 Conclusion and discussion

There is currently ample mixed-method research in longitudinal studies. Researchers use static cross-sectional mixed-method studies, but do not put them forward to examine how phenomena develop over time (Lobe & Vehovar, 2009). Bryman's (2006) literature review of mixed-method studies finds that only 47 social science studies used mix methods longitudinally in the period 1994–2003. Most of these studies coded and incorporated qualitative data in statistical analyses. Mixed-method panel surveys have the power to add much more. Method triangulation does not only provide information about the phenomena of interest from different angles, it also sheds light on methodological issues that are pertinent in longitudinal research.

One limitation of our study is the considerable initial nonresponse and some attrition in the panel survey. The initial nonresponse may have affected the results if nonrespondents differed from respondents. We believe it safe to assume that students who chose not to participate at all were less motivated than students who did participate. Including the

nonparticipants would thus have led to even more problematic factor structures and worse test results for measurement invariance.

The attrition rate between waves 1 and 2 was about 20%. Again, this may have biased the results if the students who attritted were different from those who did not. We cannot exclude this possibility entirely, but analyses of attrition based on the data from waves 1 did not show large differences, nor did we observe differences in average grades between the groups of respondents and attriters. We only conducted eight interviews. The two surveys were the core of our study, and the qualitative interviews only supplemented them. For our purposes, we think that even this limited number of interviews illustrates the type of information qualitative research methods can provide in panel surveys.

Lack of measurement invariance occurs regularly in the social sciences. We did not expect to find measurement invariance, given that the transition to university is a time of great change. Many students have to get used to a new educational system while they move out of their parental homes to live in a new city. Students who are surveyed in their first weeks of study might therefore exhibit uncertainties in more aspects than study motivations alone. During their first study year, students change in many ways that might change both the intensity and the nature of their attitudes. The assertion that attitudes, motivations, and behaviours themselves can change has deep roots in research on attitude formation (Zaller, 1992) and is acknowledged in psychological testing (for a review, see Reeve & Lam, 2005).

Nonetheless, measurement invariance, or stability of motivations and attitudes over time, is deemed necessary for comparing constructs over time. Quantitative panel studies lack the flexibility and explanatory power to explain why measurement invariance does not hold. Some attempts at constructing domain-specific scales have been made (Harter & Pike, 1984), but how to compare scores from different questionnaires over time remains unclear. We believe that mixed-method panel studies can be especially powerful when the study population undergoes major life-changing events; that is, when a 'butterfly effect' or 'gamma change' is likely. When longitudinal measurement invariance does not hold, the qualitative study can explain why. We therefore feel that mixed-method panel studies deserve more attention. Similarly, in situations where measurement invariance across sub-populations of a study does not hold, interviews might help explain how the different strata in the study population see the concept of interest differently.

We used a concurrent design so that our survey questions and qualitative interviews took place simultaneously. This design had the advantage of excluding possible differences between the survey and interview findings due to time, but it had the disadvantage of large

nonresponse in the interviewing phase. Regarding mixed-method research, Creswell and Zhang (2009) present different longitudinal designs that are applicable to research projects with exploratory purposes. In our case, we employed the qualitative part of the study to explain why measurement invariance did not hold. However, the use of both qualitative and quantitative components in panel studies can help explore how change occurs, illustrate how change affects individuals, or explain the underlying reasons for change.

7 Chapter 7: Panel Attrition

Separating stayers, sleepers and lurkers

Attrition or permanent dropout from a panel study is one of the most important sources of non-sampling error in panel surveys. Even modest attrition rates can greatly reduce the number of respondents over the course of the panel, reducing statistical power. More importantly, when attrition is selective, attrition can lead to biased survey estimates. Although the process of attrition is in many ways similar to non-response in a cross-sectional survey, there is one important difference. All respondents who drop out in a panel survey did at least participate in one wave of the study. Although panel survey managements aim to interview everyone at every wave, many respondents participate infrequently, or drop out altogether. This paper aims to show how different theoretical causes for attrition in a panel survey can be tested empirically and lead to a typology of attrition processes.

The underlying reasons that make some respondents loyal stayers and others attriters in a panel survey, can be better understood within the framework of the leverage-saliency theory (Groves, Singer, & Corning, 2000). With every request to participate in a wave of the panel survey, household members make a decision to either participate or not. Participation depends on a number of positive and negative factors (leverage) that may be of varying importance over respondents and time (saliency). The multitude of factors that either positively or negatively determine survey participation can be summarized in a propensity to participate. These factors may be stable over time (e.g. an incentive offered in every wave), but may also change (e.g. increasing respondent burden may lead to declining response propensities over time). Moreover, they also vary across respondents. Some respondents, for example, might be convinced to participate when an incentive is offered, while for others an incentive does not affect the propensity to participate at all (Laurie, Smith, & Scott, 1999). The survey methodology literature has described a number of general causes for the reason that positive and negative leverage factors of attrition vary across both individuals and time. Commitment, habit and incentives can positively affect response propensities and lead to continued participation, while panel fatigue and shocks affect the propensities negatively (Laurie, Smith & Scott 1999; Lemay 2010). We can distinguish four distinct mechanisms that can lead to declining response propensities and attrition.

The first reason for attrition is ‘absence of commitment’ (Laurie et al., 1999). Some respondents really never wanted to participate at all in the panel study, but were convinced to participate in the first wave. If

participation itself does not change their commitment to the panel survey quickly, these respondents are very likely to drop out in wave two or wave three. Conversely, when commitment is high, respondents attach value to their participation in the panel. This will result in a group of respondents who is very loyal and prolonged participation in (almost) all waves.

Repeated participation in a panel survey may lead to 'habit', even in the absence of high commitment. When survey participation becomes a habit, respondents do not longer consciously think about responding, but participate, because they have done so all along. Once this habit is broken, the respondent is subsequently at a higher risk of dropping out more often, or attriting altogether (Davidov, Yang-Hansen, Gustafsson, Schmidt, & Bamberg, 2007). Seeing panel participation as a habit explains why wave non-response in panel surveys is generally seen as an indicator for possible attrition at a later moment.

The third reason for attrition is panel fatigue. After a prolonged period of participation many respondents may feel like they have done their duty. The subjective burden that panel participation causes, weighs heavier with every wave. This leads to slowly declining response propensities until respondents drop out. The point where the burden becomes too heavy is likely to be different for every respondent (Lemay, 2010; Lipps, 2009).

The fourth reason for panel attrition is 'shock' (Lemay 2010). A shock may lead to sudden dropout from a panel. Shocks can be caused by life-changing events like a serious illness (or death), moving, or changes in the composition of the household. A shock may also be caused by one particular unpleasant experience as a panel member, like a badly designed questionnaire, the wrong use of personal data, or disturbing survey topics.

We never have direct information on the leverage and saliency factors of the decision to participate in a wave of the panel survey for both respondents and nonrespondents. Instead, attrition studies use data collected for all respondents at earlier waves, and classify respondents based on covariates that are closely related to the leverage and saliency factors.

While analyzing attrition bias on the leverage and saliency factors, some authors pool all wave-on-wave attrition patterns (Nicoletti & Peracchi, 2005; Watson & Wooden, 2009), and simply discern two groups: the attriters and stayers. This approach ignores the possibility that attrition for waves 2 to 3 is different than attrition for waves 7 to 8; it does not allow response propensities to change with time. Another

approach is to study nonresponse separately for every wave-on-wave transition (Uhrig, 2008). Apart from the fact that this yields many analyses, it is hard to deal with respondents returning to the survey, which implies that respondents can attrite multiple times. Other authors have only focused on the final state of attrition, and have limited themselves to predict whether attrition occurs or not (Tortora, 2009), or use duration models controlling for wave effects (Lipps 2009). Durrant and Goldstein (2010) take a more integrative approach and look at all possible monotone attrition patterns in a four wave panel study. With non-monotone attrition, and longer panel spans, this approach is also infeasible. Finally, Voorpostel (2009) and Behr et al (2005) follow the example of Fitzgerald et al (1998) and separate a group of attriting ('lost') from returning ('ever out') respondents, thereby also allowing for non-monotone attrition. As the panel study matures, one should however distinguish between more and more differing groups of 'ever-out' respondents.

In the data that we use in this study, respondents complete questionnaires monthly. The high frequency of data collection implies that wave non-response is even more likely to occur at any given wave than in other panel surveys, and that non-monotone attrition occurs often. The approach that we take to model attrition is different from earlier attrition studies as we model attrition in a Latent-Class framework. The underlying leverage and saliency factors that affect survey participation can be summarized in a response propensity that allows us to distinguish several classes of respondents that each follow a different attrition process. This approach allows the response propensities to vary across individuals and across time for different groups of attriters, enabling us to study who attrites, when they attrite and how the attrition process takes place. The classification involves modeling the response process with mixture-models that combine categorical and continuous latent variables. The use of Latent-Class models to study attrition has earlier been attempted by Lemay (2010), but was unsuccessful; probably due the combination of high computational demands, and the fact that ineligibility, noncontact and refusals were separately modeled.

After we discern different classes of attriters, we conclude this paper by showing how attrition classes affect attrition bias, and discuss how Latent Variable models can be successfully used to study and prevent attrition.

Who attrites?

Most of our knowledge about the correlates of attrition stems from panel studies in which respondents are interviewed by trained interviewers. In such situations, it is useful to make a distinction in attrition due to failure to locate the sample members, noncontacts, and refusals (Lepkowski & Couper, 2002). In this study, we use data from an Internet panel that contacts respondents by e-mail. We therefore do not distinguish between attrition due to nonlocation, noncontacts and refusals, but only discuss how the respondents background characteristics lead to different attrition processes within our sample. Often, there is no clear link between socio-demographic variables and attrition theories. They therefore should only lead to 'shocks' and not be important other leverage or saliency factors. They can however be important for bias assessment and correction.

Women have been shown to attrite less often than males (Behr et al., 2005; Lepkowski & Couper, 2002). Women are thought to be more conscientious and more committed and thus miss fewer waves, although evidence for this is mixed (Uhrig 2008). People with a higher Socio-Economic Status - higher education and income - attrite less, although effects are usually small (Watson and Wooden 2009). People from ethnic minorities attrite more often (Lipps, 2009). The reasons for this might be panel-specific, but we can speculate that they might perceive a higher burden due to language or cultural differences.

Other determinants of attrition are marital status (being not married), whether someone moved (or is planning to move) (Lillard & Panis, 1998) and the size of the household (Lipps, 2009). The fact that household composition is important is probably due to persuasion of other household members to stay involved in the panel survey or also drop out. Age has been found not to be related to attrition, although the oldest old and children around the age of 18 are more at risk (Lipps, 2009). Most of the effects of socio-demographic variables are either related to contactability (and thus do not apply to an Internet-panel) or seem to disappear when controlling for a change in household situation (the young), or health (the oldest old) (Jones, Koolman, & Rice, 2006).

Socio-psychological variables are deemed to have more explanatory power than demographic variables in explaining attrition and can be closely linked with attrition theories. Respondents with specific personality traits are more likely to drop out because of panel fatigue or become committed to a survey. People with high levels of agreeableness (part of the Big Five personality scale (Costa & McCrae, 1992)) are more cooperative, while conscientious people are said to be reliable, determined and have a strong need for achievement (Costa & McCrae,

1992), which should both lead to higher commitment. On the other hand, people who score high on the scale extraversion are reported to become easily bored or distracted, possibly leading to panel fatigue, drop out or infrequent response behavior (Costa & McCrae, 1992). It is not clear how neuroticism and openness to experience, the other big five personality factors, affect survey participation.

Other personality characteristics that have been linked to increased survey participation are whether people like to do cognitive tasks and evaluate. High levels of 'need for cognition' (Tuten & Bosnjak, 2001), and 'need to evaluate' (Bizer et al., 2004), should also lead to commitment to the panel survey, and prolonged survey participation.

Panel commitment and fatigue can also be measured more directly by asking respondents' attitude towards the panel survey (Rogelberg, Fisher, Maynard, Hakel, & Horvath, 2001; Stocké, 2006). Whether a respondents attributes 'value' to his own answers or 'enjoys' it indicate that commitment is present. Asking respondent directly about the burden they perceive while completing the survey, can serve as an indicator of panel fatigue (Hill & Willis, 2001), although social desirability can be a potential problem in asking the respondent directly about his survey experience.

In order to predict panel shocks, one would need detailed information on covariates in every wave; so called time-variant covariates. Practical considerations often lead panel managements to only ask about a small set of characteristics in every wave. Most often, these are related to the household composition and a few other 'core' variables, as change in address and employment situation (Uhrig 2008). One variable that is often linked to attrition, especially for older respondents is the health status at every wave (Watson & Wooden 2009), which might fluctuate with every wave. Survey methodologists have in recent years been exploring the use of paradata for explaining attrition. Similar to the respondents' attitude towards surveys, paradata can signal commitment or panel fatigue. Loosveldt, Pickery & Billiet (2002) showed that the number of item-missings is indicative of attrition in later waves. Hill & Willis (2001) furthermore hypothesize that in self-administered surveys, long interviewing time is negatively associated with future participation. Although the LISS has recorded data on all these aspects, we focus in this paper on structural or time-invariant determinants of attrition, as the inclusion of time-variant covariates further increases the high computational demands of the models we present. We will show how we can still evaluate the shock hypothesis indirectly, by studying whether response propensities in the different classes show dramatic shifts at any given time.

7.2 Methods

Data

The data for our study stem from the Longitudinal Internet Studies for the Social sciences (LISS)[14]. This panel was started in the last months of 2007, and interviews respondents monthly on a wide range of topics. The original sample for the panel was a simple random sample of Dutch households, who were contacted and recruited using a mixed-mode design. After initial contact, all household members were asked to participate in the panel survey. The participation rate in wave 1 amounted to 49 per cent (AAPOR, 2009). Those households that did not have a computer with broadband Internet connection prior to participation were provided with one (from here on called SimPC) by LISS.

For now, we only use data from the first 24 full waves of the LISS panel, spanning the period of January 2008 to December 2009. Some respondents started some months before January 2008, as the panel was built gradually. We discarded those interviews. Likewise, we chose to not include the sparse data recorded in the recruitment interview in order not to have to deal with missing data and potential mode effects. January 2008 was therefore set as the first wave of Internet-interviewing for all respondents[15]. This resulted in binary response data for 24 waves and 8148 cases. Interviewing time is about half an hour per month, and respondents receive a reward of about €15 for every hour of completing questionnaires. They are reminded in case of initial nonresponse in a specific wave, and occasionally receive information about research findings. Despite this, most panel respondents in the LISS panel fail to complete one or more of the monthly surveys, before re-entering the survey at a later time. This amounts to a total of 1983 different missing data patterns.

Instruments

We use a variety of covariates from the LISS that were mostly measured in one of the first waves of the study. Over the course of the panel, respondents in the LISS-panel were sometimes allowed to 'catch up' on questionnaires they missed at a later wave of the survey. We coded such

14 More information about the recruitment of the panel, response percentages for all waves, as well as the full questionnaires, can be found on www.lissdata.nl

15 We checked our final model results against a model where respondents were wave 1 was the actual wave 1 interview of respondents. Because of the fact that this did not alter our results, we fixed January 2008 as wave 1 for all respondents.

behavior as a wave non-response for the wave in which the questionnaire was originally fielded, but did include data on any of the covariates.

First, we use a set of socio-demographic characteristics that we treat as time-invariant: gender, age, net income (13 categories), highest education (7 categories), urbanicity, living with a partner, and having a SimPC. As psychological factors, we used the BIG-V questionnaire (Goldberg et al., 2006) to construct five factor scores (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism). Another factor score 'need to evaluate' was computed using a questionnaire by Jarvis and Patty (1996), while a factor score for 'need for cognition' was computed using the same procedure (Cacioppo & Petty, 1983). All factor scores were computed using Maximum Likelihood extraction, and oblique rotation in the case of personality. Bartlett factor scores were saved (Tabachnick & Fidell, 2007) and further used in our analyses. Other important determinants of panel attrition are the respondents' attitude towards the survey. The LISS panel contained nine questions about one's general attitude towards surveys. They ask the respondents whether they enjoy 1) internet-surveys and 2) being interviewed, whether surveys are 3) interesting and 4) important for society, whether 5) things can be learned, whether 6) completing surveys is a waste of time, 7) the perceived burden of survey requests 8) whether surveys invade privacy and 9) whether answering questions is exhaustive. A study by de Leeuw et al. (2010) has suggested that the nine items load on three factors: survey enjoyment, survey value and survey burden, but we have included all nine questions separately to assess in detail how these evaluation criteria determine panel participation.

Model

We modeled the response data using a Latent Class Framework. The advantage of using Latent Classes is that respondents are categorized based on the similarity of their response patterns. We treat a wave response in a particular wave as 1, and non-response as 0. There are three general approaches to specify the Latent Classes, that all differ in the way they treat the longitudinal nature of the data and handle unobserved heterogeneity in the data: 1) Latent Class Models (LCA), 2) Latent Class Growth Analysis (LCGA) and 3) Growth Mixture Models (GMM). In LCA all wave responses are being treated as independent from each other; i.e. the longitudinal nature of the data is ignored. Classes are formed on similar response patterns, but the response patterns in any class can take any form. LCGA explains the response patterns parametrically. Here all wave responses are explained by a Latent intercept (i), linear slope (s) and/or quadratic slope parameter (q) (see

Figure 4). This means that response patterns within every class follow a distinct pattern of growth (or here decline) in response propensities over the course of the panel study, and that this pattern is the same for everyone in this group.

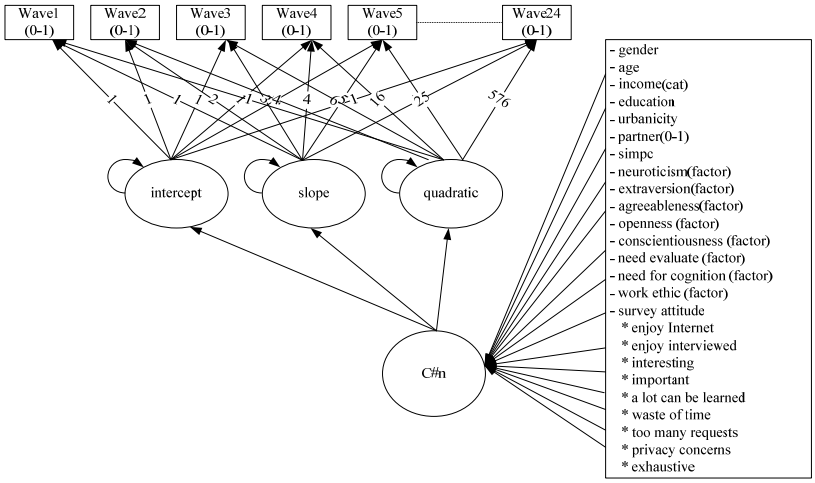

Figure 4: Latent Class Growth Analysis / Growth Mixture Model. I = intercept, s= linear slope, q=quadratic slope, C=Latent classes.[16]

The LCGA-model is less flexible than the LCA-model, but this is offset by the fact that fewer estimated parameters can lead to a better relative model fit (Kreuter & Muthen, 2008). An extension of the LCGA-model forms the Growth Mixture Model. The GMM allows for intraclass variability in the variances of the intercepts, slopes and/or the quadratic terms[17]. This allows for unobserved heterogeneity within every class (Feldman, Masyn, & Conger, 2009), which means that respondents are allowed to have a higher or lower intercept, slope and/or quadratic slope than other respondents in their class. To increase the quality of the classifications into classes, the different latent classes in all models are regressed on the set of covariates. The covariates were not regressed on the variance components of the intercept and slopes of the GMM-model, as it is not our goal to explain the variance terms of the growth parameters in the GMM-models.

[16] The figure depicts the Growth mixture model. When the variances of i, s and q are equal to 0 in the GMM, it is equivalent to Latent Class Growth Analysis.

17 In a further extension of the Growth Mixture Model, we could release the assumption that the variance terms in the GMM model would be normally distributed, leading to the estimation of a non-parametric GMM model with nodes. The estimation of such models proved to be very time-consuming and led to serious convergence problems and was therefore not pursued.

As it was unclear which of the three families of Latent Class models would explain attrition best, and how many attrition classes are necessary, we ran a set of models, each with a different number of classes, and select the model that performs best. We used three evaluation criteria for this. The Bayesian Information Criterion (BIC) serves as our primary heuristic for model comparison. This statistic is similar to absolute fit indices (for example AIC and CAIC), but it assigns a greater penalty to model complexity, and hence has a greater tendency to prefer the more parsimonious model (Nylund, Asparouhov, & Muthén, 2007). Lower values for BIC indicate a better relative fit of the model to our data. As absolute differences for BIC between competing models can be small, it is desirable to use a Bootstrapped Likelihood Ratio Test (BLRT) to specifically test whether one model fits significantly worse than the same model with one Latent Class less (Nylund et al., 2007). Because of high computational demands, we were not able to conduct such a test however. Apart from the values of BIC, we chose to also rely on values of the entropy (Celeux & Soromenho, 1996) as a criterion for the classification quality, and the substantive results of the best performing models. For this we primarily looked at observed attrition patterns of every class after estimating the model (Muthén, 2006).

All models were estimated using MPLUS 6.1 (Muthén & Muthén, 2010a). Because of the fact that individuals are clustered within households, we correct the standard errors using the robust Maximum Likelihood estimator (Muthén & Muthén, 2010b). Any missing data that we have on the covariates in our model were multiply imputed using the saturated model within the Bayesian module of Mplus 6.1 (Muthén, 2010). We initially ran all models using five imputed datasets, and repeated our analyses with twenty imputed datasets for the final model, to make sure our results were robust[18].

7.3 Results

Table 24 shows the fit of a range of tested models, each with 1 to 12 classes. In every model, we see that the BIC-values generally decrease when we add more classes, indicating that there are indeed several sub-groups in the LISS panel with a distinctive attrition pattern. When consecutively estimating the models with more classes, the BIC values reach a minimum, after which they either start increasing again. At this point, the estimation often fails to converge. Non-convergence is typical for overspecified mixture models, indicating that a more parsimonious

model should be preferred (Nylund et al., 2007). The best models in terms of their BIC values are shown in bold. The Growth Mixture Models perform best out of the three families of models. Within the group of GMM models, multiple models produce very similar BIC values. In substantive terms, these models are also largely similar, and only differ in the number of estimated classes and free variance parameters.

Table 24: BIC-values for different sets of Latent variable Mixture models explaining attrition patterns

BIC values for Model /classes	Variances of growth parameters	2	3	4	5	6	7	8	9	10	11	12
LCA	None	155125	142243	137813	135280	133883	132900	132712	131981	131839	131734	131798
LCGA	None	156277	143316	139007	136480	136838	133839	135467	132988	136033	136230	136463
GMM	i free	133558	133049	131738	131356	131179	131287	No con	132270	No con	133519	130825
GMM	s free	134942	132627	131737	131050	130951	130688	130756	130563	**130382**	**130484**	130515
GMM	q free	135901	133733	132958	132673	131666	**130774**	**130782**	130874	130741	130604	130921
GMM	is free	131437	133331	**130572**	No con	No con	No con	No con	No con	No con	No con	No con
GMM	sq free	131760	132905	**130496**	No con	No con	No con	No con	No con	No con	No con	No con
GMM	isq free	132219	No con	No con	No con	No con	No con	No con	No con	No con	No con	No con

Notes: N=8148 in all models. No con: no convergence with 150 random starts and 10 final stage iterations. The latent classes (c) were regressed on the set of covariates. BIC values shown in bold represent the best models. More information about these models is shown in Table 25.

For the six models with the lowest BIC-values (in bold), we report more model evaluation criteria in Table 25. We report the absolute fit of the model summarized in the Deviance statistic (Singer & Willett, 2003), the number of free parameters (indicating model parsimony) and the entropy.

Although several models adequately describe our data, we chose the GMM model with 4 classes and a varying slope and quadratic slope variance (GMM sq free 4) as our final model. Although this model does not have the best model fit, it has a better entropy than the GMM models with only a free slope variance. The models with only a free quadratic variance in turn have an entropy similar to the GMM sq free-4 model, but

they have a slightly worse model fit judging the values of the BIC. From this point on, we focus on the results of the GMM sq free 4 model.

Table 25: Fit statistics and model fit information of six of the best fitting models

Model	No. of classes	Deviance	BIC	No. of free parameters	Entropy
GMM s free	11	127836	130484	294	.621
GMM s free	10	127995	130382	265	.704
GMM q free	7	129171	130774	178	.740
GMM q free	8	128917	130782	207	.758
GMM is free	4	129707	130572	96	.594
GMM sq free	**4**	**129621**	**130496**	**96**	**.766**

Notes: the Deviance is calculated as -2*LogLikelihood. The entropy indicates how well the respondents can be classified (1=perfect classification, 0=totally random classification). Values higher than .8 indicate good entropy (Celeux & Soromenho, 1996). The model shown in bold is chosen as the final model

One of the primary advantages of Latent Variable models is that they allow uncertainty about model parameters. Therefore, respondents are not only assigned to one class, but class probabilities reflect the propensity to be a member of a particular attrition class. In order to show how the attrition process in the different classes takes place, we therefore discuss the response patterns for the most likely class membership of every respondent.

7.4 Attrition – when and how?

Figure 5 shows the observed posterior response probabilities for each class[19]. The first group of respondents in the panel is comprised of a group of respondents who almost always participate in the panel survey. We call these 'loyal stayers'. While this group makes up about 12 per cent of the panel, the largest group consists of a group of 65 per cent, whom we call 'gradual attriters'. These are respondents who participate in most waves, but do occasionally miss out on one or more waves. This group of respondents shows response propensities around .9 at the start of the panel. Their response propensities do decline somewhat over time, but not very fast. At the end of the 24 waves in our analyses, they still have propensities around .6. In the mean time, response propensities vary over

[19] The posterior probabilities were derived by running the final model on only 1 imputed dataset, while fixing all parameters of this model to the solution which was found with 20 imputations. The most likely class membership was then used to plot the posterior response probabilities.

the waves, following a slow downward trend. An exception to the slow downward trend is the abrupt decline in response propensities at wave 6.

The third class (7 per cent) consists of respondents who we will label 'lurkers'. This label represents the fact that respondents in this class participate very infrequently. Their response propensities in the first waves are very low, but they do increase to about .5 around wave 9, meaning that respondents in this class take part in about every second survey. After wave 9, the lurkers seem to mimic the declining response propensities of the group of gradual attriters, albeit at a lower level. The final class of respondents (17 per cent) follows a typical pattern of 'fast attrition'. These respondents start out with high response propensities around .9, but propensities then quickly decline to about .3 in wave 7. By wave 20, these respondents have all dropped out of the study.

The top part of Table 26 represents the fitted growth curves in every class. The variances of the slope (s) and quadratic slope (q) allow individual variance within every class. The variances in the group of attriters and loyal stayers are quite small, but they are larger in the groups of lurkers and gradual attriters (see Appendix D).

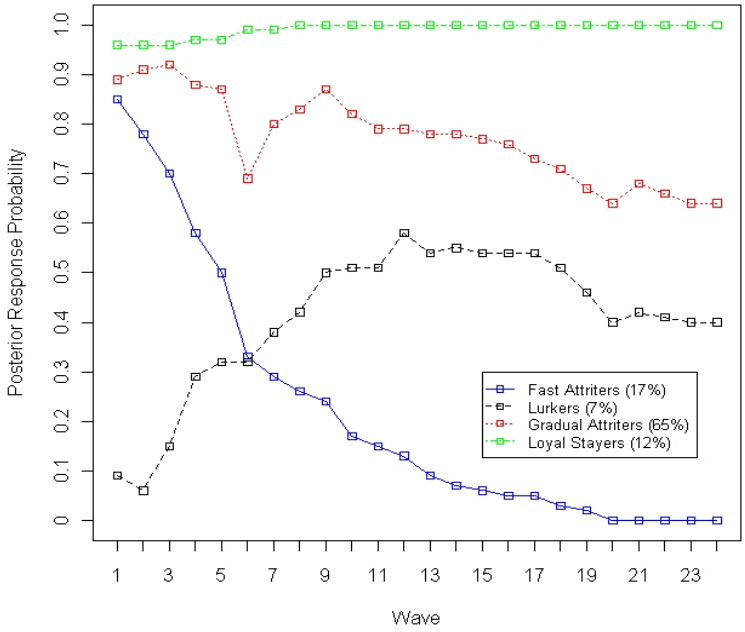

Figure 5: Posterior wave response probabilities and sizes for the most likely class membership for the Growth Mixture Model with a varying slope (s) and quadratic slope (q) variance within the 4 classes

Table 26: Unstandardized growth parameters and multinomial regression coefficients (logit) of the covariates (X) on class membership (C). Standard errors shown in brackets.

Growth Parameters/ Class	1 loyal stayers	2 gradual attriters	3 lurkers	4 fast attriters
i	-	108.34 (4.19)	-5.01 (.58)	-.42 (.52)
s	.62 (.11)	537.47 (240.47)	5.33 (1.27)	6.00 (2.23)
q	-.43 (.04)	998.65 (807.48)	-1.98 (.39)	-16.71 (2.82
var s	10.45 (.95)	6672.83 (8459.17)	8.08 (.86)	79.56 (17.48)
var q	1.25 (.11)	4542.43 (7467.56)	3.23 (.64)	86.77 (35.66)
cov(s,q)	-3.16 (.32)	-3.16 (.32)	-3.16 (.32)	-3.16 (.32)
Logit coefficients/ Covariates				
Socio-demographic				
Gender (1=f)	-	.12 (.10)	-.29 (.15)*	.11 (.12)
Age	-	-.04 (.01)**	-.07 .01)**	-.02 (.01)*
Income (13 cat)	-	.02 (.02)	-.05 (.05)	.04 (.04)
Education (7 cat)	-	.01 (.05)	-.22(.07)**	-.09 (.05)
Urbanicity	-	-.17 (.10)	-.17 (.08)*	-.15 (.06)**
Partner(1=yes)	-	.03 (.13)	.07 (.16)	.21 (.19)
SimPC (1=yes)	-	-.47 (.29)*	-.88 .42)**	-1.24 (.35)**
Psychological				
Openness	-	.22 (.07)**	.18 (.08)*	.20 (.07)**
Conscientiousness	-	-.35 (.07)**	-.27 .08)**	-.29 (.07)**
Extraversion	-	.29 (.07)**	.17 (.08)*	.28 (.06)**
Agreeableness	-	.17 (.07)*	.33 (.09)**	.21 (.07)**
Neuroticism	-	-.02 (.08)	-.01 (.08)	.00 (.07)
Need to evaluate	-	-.07 (.06)	-.07 (.08)	.00 (.07)
Need for cognition	-	.01 (.04)	.10 (.09)	-.07 (.06)
Survey attitude	-			
-1 Enjoy Internet	-	-.23 (.08)**	-.27 .07)**	-.39 (.05)**
-2 Enjoy -interviewed	-	-.01 (.07)	.36 (.11)**	-.07 (.04)
-3. Interesting	-	-.08 (.04)*	-.03 (.09)	-.15 (.06)**
-4. Important	-	.00 (.06)	.09 (.09)	.04 (.07)
-5. A lot can be learned	-	.03 (.06)	-.08 (.09)	.06 (.06)
-6. Waste of time	-	.15 (.04)**	.18 (.07)**	.18 (.06)**
-7. Too many requests	-	-.02 (.03)	-.03 (.05)	-.04 (.04)
-8. Privacy concerns	-	.00 (.03)	.07 (.05)	.04 (.05)
-9. exhaustive	-	.14 (.04)**	.17 (.05)**	.19 (.04)**

Notes: N=8148. The intercept in the first class is not estimated in order to estimate the intercept parameters in the other classes. Var q: variance of the slope parameter in every class. Var q: variance of the quadratic slope in every class. Cov(s,q): covariance of s and q parameters.

The reported coefficients are unstandardized, and represent multinomial regression coefficients using class 1 ("loyal stayers") as the reference class.

* p<.05, ** p<.01

7.5 The characteristics of attriters

We will now describe how the four attrition classes differ on the covariates which were used predict respondent classification. The coefficients shown in the lower part of Table 26 represent the logit values of being in classes 2 to 4 versus being in class 1. We favor the use of logit parameters over odds-ratios to be able to compare the predictive power of every covariate[20].

First, we look at the socio-demographic predictors. Surprisingly, we do not find that males attrite more often than women. The logit parameter of .12 for example means that compared to the class of loyal stayers, we find the proportion of females to be .12 log-odds higher in the class of gradual attriters. This translates into an odds ratio of 1.12, holding all other variables constant. Compared to the class of stayers, we only find more males in the class of lurkers. In all attriting classes, we find younger people compared to the loyal stayers, and we find the largest effect in the class of lurkers. Respondents in this class are also significantly less educated than those in the other classes, and live in slightly more urbanized areas. Fast attriters are also found more in the urban areas. Furthermore, we find that the attriting classes have received a SimPC less often than the class of stayers. The effects of having a SimPC provided are large. Almost none of the people who received a SimPC are to be found in the attriting classes.

With the psychological variables we can more directly evaluate whether we find the attrition processes to correspond with theories on the causes of attrition. The psychological variables separate the stayers from the gradual and fast attriters on the one hand, and the lurkers on the other hand. We find people in all three classes of attriters to be less conscientious, more open and extravert, and the class of lurkers to be more agreeable and less extravert.

Almost all these findings are in line with the commitment hypothesis. All attriters are less conscientious and more agreeable than the stayers which implies that attriters are less committed to being a good panel respondent. We unexpectedly find the attriters to also be more open to new experiences. This may imply that attritting people were lured into panel participation, without really being motivated, and that once in the panel, they find participating not exciting.

Together with the psychological variables, the survey attitude variables have the largest logit parameters, thus explaining the differences between the attrition classes best. We find consistent differences between the class of loyal stayers and the three attriting classes. All three

[20] Odds ratio's can easily be calculated by taking the exponent of the logit (log-odds) parameter estimates.

groups of attriting respondents find the fact that they participate in an Internet survey more a waste of time and more exhaustive. We however also find three differences between the attriting classes of respondents. First, fast attriters enjoy Internet surveys even less than the other attrition groups. Second, lurkers do report that they dislike Internet surveys, but do enjoy being interviewed in person, implying that they specifically dislike the fact that the LISS is an Internet panel survey. Third the fast attriters find surveys less interesting than the gradual attriters and lurkers.

In summary, the loyal stayers are more conscientious, less open extravert, agreeable, older, live in more rural areas and enjoy the survey more. The lurkers stand out as being a lot younger and less educated, but they report that they do enjoy completing surveys. The fast attriters stand out as experiencing the greatest burden and the least commitment. Overall, we find evidence that lack of commitment and panel fatigue do explain attrition for the different classes, and that the differences in attrition patterns can be explained by differences in levels of commitment and panel fatigue.

Apart from commitment and panel fatigue, habit and shock are the other hypothetical causes of attrition. Although we cannot evaluate these causes directly, the response propensities shown in Figure 5 can be used to evaluate these causes indirectly. The response propensities in the classes of gradual and fast attriters show a sudden decline at wave 6 of the panel survey. Wave 6 was fielded in June 2008, before the summer holidays, so the timing of this wave is not the cause. We believe this shock to have occurred because of the topic of that month's survey. Respondents had to report in detail about their household's sources of income, details of their income, as well as their expenses. The interviewing time for this topic amounted to about half an hour, which obviously caused many respondents not to start that wave, or break it off halfway. Response propensities for the fast and gradual attriters drop by about 0.2 in wave 6. In the class of gradual attriters, respondents return to the panel after wave 6, so that their response propensities are back at 0.9 at wave 11. In the group of fast attriters, respondents are however permanently lost after wave 6. To evaluate the shock hypothesis more formally, we would need time-varying covariates on for example household situation, moving and health status, which is beyond the scope of this paper. The response propensities in Figure 5 do show however that boring questionnaires can either lead to direct attrition (shock), or can break the habit of responding to survey requests, starting or accelerating a downward trend in response propensities leading to attrition.

7.6 Attrition – does it matter?

Apart from looking at the characteristics of those who attrite, it is also interesting to see how attrition matters for substantive statistics. Here we focus on how the different attrition classes contribute to attrition bias in the estimate of the Dutch parliamentary election results in 2006. We chose the variable voting behavior on purpose, as we can validate the survey estimates using all respondents in the panel, the respondents in the various attrition classes and those who remain in the panel at wave 24. Voting behavior was recorded twice in the first waves of the panel, so we have information for most of the panel members[21]. Table 27 shows the actual election results (in percentages) for the general election in the second column. The third and fourth columns show the results for the respondents in the LISS panel. We see that considerable bias already exists at the start of the panel, most likely due to nonresponse in the panel recruitment phase, although some measurement errors should not be excluded as a potential cause. The fifth to eighth column show estimates for the election results in every class, as well as the absolute difference (summed) with the official election result and the relative contribution of every class to attrition bias. At the start of the panel, the absolute bias adds up to 10 percentage points when compared to the official result. 23 months later, using only the people who are still active at the end of our study, attrition bias has decreased to 9.3 percentage points. All classes contribute to this bias, although we see that the bias is large in the group of loyal stayers (16 per cent) and especially the lurkers (34 per cent). Table 27 shows for example that 32.3 per cent of the stayers, and 18.8 per cent of the lurkers report voting for the Christian democrats whereas in the real election, 26.5 per cent of the Dutch electorate did. Because of the fact that the class of lurkers is small, we find the relative contribution of this class to attrition bias for all parties to be 17 per cent. The majority of the attrition bias stems from the class of gradual attriters, as that is the largest class in our study.

[21] We only have missing information for about 50% of the group of lurkers. This is because of higher levels of nonresponse, and a higher proportion of respondents in this class that were ineligible to vote at the time of the parliamentary election. We cannot exclude the possibility that the exclusion of these people introduces new bias to our results, but exploratory analyses on other variables found no differences between these the groups of lurkers for whom we have data on their voting behavior, and those for whom we have not.

Table 27: Estimates of the election result and the contribution to Nonresponse bias of every attrition class

Party/voting percentages	Election results	wave1	wave24	1 Fast	2 gradual	3 Lurkers	4 Stayers
Chr. Dem. (CDA)	26.5	25.1	26.0	22.4	24.2	18.8	32.3
Labour (PvdA)	21.2	19.3	19.5	19.0	19.5	16.2	19.4
Socialists (SP)	16.6	17.4	17.6	17.0	17.4	23.8	16.6
Liberals (VVD)	14.7	15.8	15.0	17.4	15.6	18.8	14.7
Freedom (PVV)	5.9	4.3	3.7	4.6	4.4	3.8	3.4
Green Left (GL)	4.6	6.2	6.2	8.6	6.6	2.5	2.6
Chr. Union (CU)	4.0	4.9	5.3	4.4	4.9	5.0	5.3
Others	6.5	7.2	7.2	6.6	7.4	11.1	5.7
Absolute bias	-	10.0	9.3	15.2	11.0	33.8	15.6
Percentage contribution	-	-	-	18.6	50.8	17.1	13.5

Notes: N=5013. This includes all panel members who responded in January 2008 (wave 1).

The names of the political parties are translated and abbreviated. For the original names, see www.lissdata.nl. Column 2 denotes the true parliamentary results. Column 3 represents the election result estimated with the data as provided by respondents in wave 1, thus only including initial nonresponse bias. The fourth column denotes the same estimate, but only using the respondents who responded to wave 24. The fifth to eighth columns, denote the estimate of the election results, using only respondents in classes 1 to 4. These results are unweighted for initial sample selection and nonresponse in the panel recruitment phase.

The absolute bias indicates the bias in each class or wave as compared to the official election result, unweighted for class size (sum of all party biases). The percentage contribution shows the contribution of each class in the total absolute Nonresponse bias, weighted for the size of each class.

7.7 Conclusion and Discussion

This paper showed how attrition can be described as a process that varies over individuals and time. The underlying leverage and saliency factors that affect survey participation can be summarized in a response propensity that allows us to distinguish several classes of respondents that follow a different attrition process. The analysis model that we propose corresponds to substantive theories about attrition, and overcomes analytical problems in previous attrition studies. In the context of a panel survey, the leverage-saliency theory posits that the decision to participate in a (wave of a) panel survey is determined by positive and

negative factors (leverage), that can increase or decrease in importance (saliency) with time and across respondent. Almost all of the respondents in our study miss on or more waves of the study. Sometimes, wave-non response leads to permanent drop-out, but more often, respondents return to the panel survey. The group of 'ever out' respondents is diverse and consists of stayers, gradual and fast attriters and lurkers. These groups differ from each other not only in their response patterns, but also on substantive variables. Attriters have a different type of personality and value survey participation differently from loyal stayers, which leads to differences in the levels of leverage and saliency factors for these respondents and different attrition patterns. We only have proxy information on the leverage and saliency factors, but our results suggest that attriters have less commitment, and higher levels of panel fatigue.

We were not able to directly test other theoretical causes of attrition – shock and habit; we would need time-variant covariates to assess how variables such as health status, household situation or unpleasant panel experience might affect the leverage and saliency factors at every wave separately. The inclusion of time-variant covariates should not fundamentally alter the Latent Class model we used. One would expect the time-variant covariates to have no effect on responses in the groups of loyal stayers, nor in the group of fast attriters. However, they should strongly predict attrition for the class of slow attriters, who show greater panel fatigue in our study. For them, a shock, whether it is in the form of a life-event or an unpleasant panel experience, can make the balance of positive and negative survey participation factors tip firmly to the negative and lead to attrition. The fact that we observe a large decline in response propensities for this group in the waves with the long income questionnaires is a clear sign of this. This finding shows how questionnaire design and the survey process itself are very important in the attrition process. Although we can only indirectly test this one example of a 'shock' effect, it would be interesting to see whether including time-varying covariates increase the importance of shocks for panel attrition in specific classes. Estimation of Growth Mixture models is time-consuming, and adding time-covariates increases estimation time substantially. Future increases in computing power should solve this problem.

Further analyses into attrition processes should not only focus on attrition errors, but take all survey errors into account. In this paper, we explicitly chose not to study any survey errors that were introduced prior to the start of the panel. Although we want to stress the importance of the panel composition stage for limiting the size of total survey error, we here focused on the determinants of non-response conditional on enrolment in the survey. Ideally, any study of attrition, should not only

study errors because of initial non-response and attrition, but also measurement errors. Panel managements could try to prevent or limit attrition and initial non-response, but if this comes at the price of lower data quality, pursuing tailoring strategies may come at a price of decreasing data quality. Research in cross-sectional surveys has suggested that more reluctant respondents also have the lowest data quality (Tourangeau, Groves, & Redline, 2010). One way forward to incorporate measurement errors in attrition models is to include indicators of the response quality per class.

We only tested attrition bias for voting behavior, and the possibility remains that non-response bias is different for other variables. For the LISS panel, it seems that bias was introduced in the panel composition phase, and that attrition does not make this bias much worse. However, attrition bias is large in the class of loyal stayers, who would logically comprise an increasingly large proportion of the panel, possibly leading to more bias with time. It is therefore important to try and keep the classes at risk of attriting in the panel.

The final question that remains unanswered is whether attrited respondents in the LISS panel have really dropped out forever. Many respondents miss out on one or more waves towards the end of our study. The fact that we find only seventeen per cent of respondents to have dropped out altogether is hopeful however. In 2010, a project was started to see if and under what circumstances respondents wanted to return. In a future study, we plan to see if and for how long respondents from the classes of gradual attriters, lurkers and fast attriters can be turned into loyal stayers.

8 Conclusion

The five studies in this dissertation all highlight aspects of survey data quality in panel surveys. The goal of each study is to show how survey errors can be studies, and how they affect survey estimates. In all studies, statistical models are used to study how panel design features affect survey errors. The main message of this dissertation is that survey errors can be studies in panel surveys, and that decisions in the design of panel surveys can be evaluated. The studies in this dissertation should help future studies to evaluate survey errors more effectively.

Many topics or issues in panel survey design were not discussed. Questionnaire design and fieldwork procedures are among the topics that are very important, but not discussed. This does not mean these topics are less important than the topics discussed here. Rather to the contrary, day-to-day survey operations in trying to keep in touch with respondents, is very important for limiting attrition rates, and keeping respondents motivated. Motivated respondents in turn are good respondents: they will try to provide data with high quality. Similarly, designing good survey questions does not only lead to less measurement error, but also motivated respondents.

This dissertation showed how survey design choices may affect different components of survey error, and ways in which possible trade-offs between these components can be studied further. The paper in chapter 7 showed that attrition in a panel survey can take many different forms and shapes. Some respondents are loyal respondents, and participate in every wave, others drop-out, while again other respondents participate infrequently. For each group of respondents, measurement errors can be studied by using re-interview data or validation data. Re-interview data can be modeled using a quasi-simplex model. If re-interview data for attriting respondents are unavailable, factor models or validation data can be used to also gauge how different attrition classes contribute to measurement error. Chapter 7 showed how the different classes of attriters report very different voting behaviors, and it is likely that they also differ on other aspects. Social scientists worry about the consequences of attrition on substantive survey estimates. A study that would study measurement errors for each class of attriters can answer whether these worries are justified. The same study could also study how problematic measurement errors are for the class of loyal respondents. There is belief that ongoing participation in a panel survey may lead to 'panel conditioning'. Because of participation in a panel survey, the attitudes or behaviors of respondents themselves may be subject to change.

Panel conditioning is notoriously difficult to study, because any conditioning effects interact with attrition. Although a Latent Variable Model like the one presented in chapter 7 may shed some light on how different the group of loyal respondents behaves over time, a propensity matching procedure as explored in chapter 3 may prove a right alternative. This approach has been explored by others (see Kasprzyk et al.1989 for example) but matching may improve the ability to compare equivalent groups.

Studying examples like these should teach us how different aspects of survey error interact. Understanding total survey error within the context of panel surveys should inform panel survey methodologists how to set up panel surveys. Should we use mixed-mode surveys to decrease nonresponse error at the risk of introducing more (and different) measurement errors? Should we try to limit attrition at all costs, or devote our time to try and keep specific respondents in the panel survey?

The data quality of change estimates can be improved by using Dependent Interviewing and edit checks, as chapters 4 and 5 showed. The effect of introducing both has a limited effect on substantive estimates that are derived from the income questions for which Dependent Interviewing and edit checks are used. Even if we understand how survey errors interact, it remains a question how this matters for substantive estimates. This is a question that every methodological study should consider.

9 References

AAPOR. (2008). *Standard definitions: Final dispositions of case codes and outcome reports for surveys, 5th edition*. Lenexa, Kansas: American Association for Public Opinion Research.

AAPOR. (2009). *Standard definitions and eligibility calculation.* Lenexa, Kansas: American Association for Public Opinion Research.

Alwin, D. F. (2007). *Margins of error - A study of reliability in survey measurement*. Hoboken, New Jersey: Wiley.

Arbuckle, J. L. (2007). *AMOS 16.0.1*. Spring House, Pensyllvania: AMOS development corporation.

Behr, A., Bellgardt, E., & Rendtel, U. (2005). Extent and determinants of panel attrition in the European community household panel. *European Sociological Review, 21*(5), 489-512.

Bethlehem, J. G., & Keller, W. J. (1987). Linear weighting of sample survey data. *Journal of Official Statistics, 3*(2), 141-153.

Biemer, P. P. (2003). *Introduction to survey quality*. Hoboken, NJ: John Wiley Sons, Inc.

Biemer, P. P. (2010). Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly, 74*(5), 817.

Biemer, P. P. (2001). Nonresponse bias and measurement bias in a comparison of face to face and telephone interviewing. *Journal of Official Statistics, 17*(2), 295-320.

Bizer, G. Y., Krosnick, J. A., Holbrook, A. L., Wheeler, S. C., Rucker, D. D., & Petty, R. E. (2004). The impact of personality on cognitive, behavorial, and affective political processes: The effects of need to evaluate. *Journal of Personality, 72*(5), 995-1028.

Blyth, B. (2008). Mixed mode: The only 'fitness' regime? *International Journal of Market Research, 50*(2), 241-266.

Boeije, H. R. (2010). *Analysis in qualitative research*. London: Sage Publication Ltd.

Böheim, R., & Jenkins, S. P. (2006). A comparison of current and annual measures of income in the British household panel study. *Journal of Official Statistics, 22*(4), 733-758.

Bryman, A. (2006). Integrating quantitative and qualitative research: How is it done? *Qualitative Research, 6*(1), 97-113.

Cacioppo, J. T., & Petty, R. E. (1983). Effects of need for cognition on message evaluation recall and persuasion. *Journal of Personality and Social Psychology, 45*(4), 805-818.

Callegaro, M. (2008). Seam effects in longitudinal surveys. *Journal of Official Statistics, 24*(3), 387-403.

Campbell, D. T. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81.

Celeux, G., & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification, 13*, 195-212.

Conrad, F. C., Rips, L. J., & Fricker, S. S. (2009). Seam effects in quantitative responses. *Journal of Official Statistics, 25*(3), 339-361.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management, 27*(4), 724-750.

Costa, Jr., P. T., & McCrae, R. R. (1992). Normal personality assessment in clinical practice: The NEO personality inventory. *Psychological Assessment, 4*(1), 5-13.

Creswell, J. W., & Plano Clark, V. L. (2007). *Designing and conducting mixed methods research*. London: Sage.

Creswell, J. W., & Zhang, W. (2009). The application of mixed methods design to trauma research. *Journal of Traumatic Stress, 22*(6), 612-621.

Davidov, E., Yang-Hansen, K., Gustafsson, J., Schmidt, P., & Bamberg, S. (2007). Does money matter? A theory-driven growth mixture model to explain travel-mode choice with experimental data. *Methodology, 2*(3), 124-134.

de Leeuw, E. D. (2005). To mix or not to mix data collection modes in surveys. *Journal of Official Statistics, 21*(2), 233-255.

de Leeuw, E. D., & de Heer, W. (2002). Trends in household survey nonresponse: A longitudinal and international comparison. In R. M. Groves, D. A. Dillman, J.

L. Eltinge & R. J. A. Little (Eds.), *Survey nonresponse* (pp. 41-54). New York: Wiley.

de Leeuw, E. D., Hox, J. J., Lugtig, P., Vis, C., Göritz, A., Bartsch, S., et al. (2010). Does familiarity breed contempt? Measuring and comparing survey attitude among new and repeat respondents cross-culturally. Paper presented at the 63rd WAPOR conference, Chicago, 13 May 2010.

de Leeuw, E. D., & van der Zouwen, J. (1989). Data quality in telephone and face to face surveys: A comparative meta-analysis. In R. M. Groves, P. B. Biemer, L. E. Lyberg, J. T. Massey, W. L. I. Nichols & J. Waksberg (Eds.), *Telephone survey methodology* (pp. 283-299). New York: Wiley.

Deheji, R. H., & Wahba, S. (2002). Propensity score matching methods for non-experimental causal studies. *The Review of Economics and Statistics, 84*(1), 151-161.

Department for Work and Pensions. (2008). *Households below average income 2005/06* No. 2009). London: Department for work and pensions

Dillman, D. A. (2007). *Mail and internet surveys - the tailored design method* (2nd edition ed.). Hoboken, New Jersey: John Wiley & Sons.

Dillman, D. A., & Christian, L. M. (2005). Survey mode as a source of instability in responses across surveys. *Field Methods, 17*(1), 30-52.

Dillman, D. A., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J., et al. (2009). Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (IVR) and the internet. *Social Science Research, 38*(1), 1-18.

Durrant, G. B., & Goldstein, H. (2010) Analysing the probability of attrition in a longitudinal survey. *University of Southampton Working Paper,* (M10/08)

Eid, M. (2009). The multitrait-multimethod matrix at 50!. *Methodology, 5*(3), 71.

Feldman, B. J., Masyn, K. E., & Conger, R. D. (2009). New approaches to studying problem behaviors: A comparison of methods for modeling longitudinal, categorical adolescent drinking data. *Developmental Psychology, 45*(3), 652-676.

Fitzgerald, J., Gottschalk, P., & Moffitt, R. (1998). An analysis of sample attrition in panel data: The Michigan panel study of income dynamics. *Journal of Human Resources, 33*, 251-299.

Galambos, N. L., Almeida, D. M., & Petersen, A. M. (1990). Masculinity, femininity, and sex role attitudes in early adolescence: Exploring gender intensification. *Child Development, 61*(6), 1905-1914.

Glasner, T. (2011). *Reconstructing event histories in standardized survey research: Cognitive mechanisms and aided recall techniques.* Unpublished Free University, Amsterdam.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., et al. (2006). The international personality item pool and the future of public domain personality measures. *Journal of Research in Personality, 40*, 84-96.

Golembiewski, R. T., Billingsley, K., & Yeager, S. (1976). Measuring change and persistence in human affairs: Types of change generated by OD designs. *Journal of Applied Behavorial Science,* (12), 133-157.

Groves, R. M. (1978). An empirical comparison of two telephone sample designs. *Journal of Marketing Research, 15*, 622-631.

Groves, R. M. (2005). *Survey errors and survey costs* (2nd ed.). Hoboken, New Jersey: John Wiley & Sons.

Groves, R. M., & Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias - A meta analysis. *Public Opinion Quarterly, 72*(2), 167-189.

Groves, R. M., Singer, E., & Corning, A. (2000). Leverage-saliency theory of survey participation - description and an illustration. *Public Opinion Quarterly, 64*, 299-308.

Groves, R. M., Fowler jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. Hoboken: Wiley & Sons.

Hale, A., & Michaud, S. (1995). Dependent interviewing: Impact on recall and on labour market transitions. *SLID Research Paper, 1995-06*

Harter, S., & Pike, R. (1984). The pictorial scale of perceived competence and social acceptance for young children. *Child Development, 55*, 1969-1982.

Haughton, D. M. A., Oud, J. H. L., & Jansen, R. A. R. G. (1997). Information and other criteria in structural equation model selection. *Communications in Statistics - Simulation and Computation, 26*(4), 1477-1516.

Heerwegh, D. (2009). Mode differences between face-to-face and web surveys: An experimental investigation of data quality and social desirability effects. *International Journal of Public Opinion Research, 21*(1), 111-121.

Heise, D. R. (1969). Separating reliability and stability in test-retest correlation. *American Sociological Review, 34*, 93-101.

Hidi, S., & Harackiewicz, J. M. (2000). Motivating the academically unmotivated: A critical issue for the 21st century. *Review of Educational Research, 70*(2), 151-179.

Hill, D. H. (1994). The relative validity of dependent and independent data collection in a panel survey. *Journal of Official Statistics, 10*(4), 359-380.

Hill, D. H., & Willis, R. J. (2001). Reducing panel attrition: A search for effective policy instruments. *The Journal of Human Resources, 36*(3), 416-438.

Ho, D. E., Stuart, E., Imai, K., & King, G. (2009). *Package 'matchit'* , last visisted on the 23rd of November 2009, available on http://cran.r.project.org/web/packages/Matchit/Matchit.pdf

Holmberg, A. (2004). Pre-printing effects in official statistics: An experimental study. *Journal of Official Statistics, 20*(2), 219-232.

Hoogendoorn, A. W. (2002). Evaluation of a questionnaire design for dependent interviewing in a web survey. Paper presented at the *International Conference on Questionnaire Development, Evaluation and Testing Methods,* Charleston, South Carolina.

Hoogendoorn, A. W. (2004). A questionnaire design for dependent interviewing that addresses the problem of cognitive satisficing. *Journal of Official Statistics, 20*(2), 219-232.

Iacus, S. M., King, G., & Porro, G. (2009). *Matching for causal inference without balance checking* , last visited on August 30th, 2009. available on http://gking.harvard.edu/files/abs/cem-abs.shtml

Jäckle, A. (2008). Dependent interviewing: Effects on respondent burden and efficiency in data collection. *Journal of Official Statistics, 24*(3), 411-430.

Jäckle, A. (2008). Measurement error and data collection methods: Effects on estimates from event history data. *ISER Working Paper, 2008-13*

Jäckle, A. (2009). Dependent interviewing: A framework and application to current research. In P. Lynn (Ed.), *Methodology of longitudinal surveys* (pp. 93-112). Chichester: Wiley.

Jäckle, A., Laurie, H., & Uhrig, S. N. (2007). The introduction of dependent interviewing on the British household panel survey. *ISER Working Paper, 2007-7*

Jäckle, A., & Lynn, P. (2007). Dependent interviewing and seam effects in work history data. *Journal of Official Statistics, 23*(4), 529-551.

Jäckle, A., Sala, E., Jenkins, S. P., & Lynn, P. (2004). Validation of survey data and employment: The ISMIE experience. *ISER Working Paper, 2004-14*

Jarvis, S., & Jenkins, S. P. (2000). Low-income dynamics in 1990s Britain. In D. Rose (Ed.), *Researching social and economic change* (pp. 188-210). London: Routledge.

Jenkins, S. P. (2008). Marital splits and income changes over the longer term. In M. Brynin, & J. Ermisch (Eds.), *Changing relationships* (pp. 217-236). Abingdon: Routledge.

Jones, A. M., Koolman, X., & Rice, N. (2006). Health-related non-response in the British household panel survey and European community household panel: Using inverse-probability-weighted estimators in non-linear models. *Journal of the Royal Statistical Society Series A, 169*(3), 543-569.

Jorgensen, D. L. (1989). *Participant observation. A methodology for human studies*. Newbury Park: Sage.

Kaminska, O. (2010). Satisficing among reluctant respondents in a cross-national context. *Public Opinion Quarterly, 74*(5), 956.

Kasprzyk, D., Duncan, G. J., Kalton, G., & Singh, M. P. (1989). *Panel surveys*. New York: Wiley.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (second ed.). New York: The Guildford Press.

Kool, L., Maris, A., & Munck, S. D. (2009). *Marktrapportage elektronische communicatie (market report on electronic communication)*, Netherlands Organisation for the Advancement of Science (TNO)

Kreuter, F., & Muthen, B. (2008). Longitudinal modeling of population heterogeneity: Methodological challenges to the analysis of empirically

derived criminal trajectory profiles. In C. R. Hancock, & K. M. Samuelson (Eds.), *Advances in latent variable mixture models* (pp. 53-75). Charlotte NC: Information Age Publishing.

Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social desirability bias in CATI, IVR and web surveys. *Public Opinion Quarterly, 72*(5), 847-865.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5*, 213-236.

Kuckartz, U. (2007). *MaxQDA - professional software for qualitative data analysis*. Berlin: Verbi GMBH

Laurie, H., Smith, R., & Scott, L. (1999). Strategies for reducing nonresponse in a longitudinal panel survey. *Journal of Official Statistics, 15*(2), 269-282.

Lee, S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel web surveys. *Journal of Official Statistics, 22*(2), 329-349.

Lemay, M. (2010). *Understanding the mechanism of panel attrition.* Unpublished Doctoral thesis, University of Maryland

Lepkowski, J. M., & Couper, M. P. (2002). Nonresponse in the second wave of longitudinal household surveys. In Groves, Robert M. et al. (Ed.), *Survey nonresponse*. New York: John Wiley & sons.

Lillard, L. A., & Panis, C. W. A. (1998). Panel attrition from the panel study of income dynamics - household income, marital status and mortality. *The Journal of Human Resources, 33*(2), 437-457.

Lipps, O. (2007). Attrition in the Swiss household panel. *Methoden - Daten - Analyses, 1*(1), 45-68.

Lobe, B., & Vehovar, V. (2009). Towards a flexible online mixed method design with a feedback loop. *Quality and Quantity, 43*(585-597)

Loosveldt, G., Pickery, J., & Billiet, J. (2002). Item nonresponse as a predictor of unit nonresponse in a panel survey. *Journal of Official Statistics, 18*, 545-557.

Lynn, P. (Ed.). (2009). *Methodology of longitudinal surveys*. Chichester: Wiley.

Lynn, P. (2009). Methods for longitudinal surveys. In P. Lynn (Ed.), *Methodology of Longitudinal surveys* (pp. 1-20). Chichester: Wiley.

Lynn, P. (2011). Maintaining cross-sectional representativeness in a longitudinal general population survey. *Understanding Society Working Paper Series, 2011-04*

Lynn, P., Buck, N., Burton, J., Jäckle, A., & Laurie, H. (2005). A review of methodological research pertinent to longitudinal survey design and data collection. *ISER Working Paper, 2005-29*

Lynn, P., Buck, N., Burton, J., Laurie, H., & Uhrig, S. C. N. (2006). *Quality profile: British household panel survey version 2.0: Waves 1 to 13: 1991-2003*. Colchester; Colchester: Institute for Social and Economic Research, University of Essex; Institute for Social & Economic Research.

Lynn, P., Jäckle, A., Jenkins, S. P., & Sala, E. (2006). The effects of dependent interviewing on responses to questions on income sources. *Journal of Official Statistics, 22*(3), 357-384.

Lynn, P., Jäckle, A., Jenkins, S. P., & Sala, E. (2012, forthcoming). The impact of interviewing method on measurement error in panel survey measures of benefit receipt: Evidence from a validation study. *Journal of the Royal Statistical Society A,*

Marquis, K. H., Marquis, M. S., & Polich, J. M. (1986). Response bias and reliability in sensitive topic surveys. *Journal of the American Statistical Association, 81*(394), 381-389.

Mason, J. (2006). Mixing methods in a qualitative driven way. *Qualitative Research, 6*(1), 9-25.

Mathiowetz, N. A., & McGonagle, K. A. (2000). An assessment of the current state of dependent interviewing in household surveys. *Journal of Official Statistics, 16*(4), 401-418.

Moore, J. (2006). The effects of questionnaire design changes on general income amount nonresponse in waves 1 and 2 of the 2004 SIPP panel. *Research Report Series - Survey Methodology,* 2006-4

Moore, J., Bates, N., Pascale, J., Griffiths, J. K., & Okon, A. (2006). Use of dependent interviewing procedures to improve data quality in the measurement of change. *Research Report Series - Survey Methodology, 2006-2*

Moore, J., Bates, N., Pascale, J., & Okon, A. (2009). Tackling seam bias through questionnaire design. In P. Lynn (Ed.), *Methodology of longitudinal surveys* (pp. 73-92). Chichester: Wiley.

Moore, J., Stinson, L. L., & Welniak, E. J. (2000). Income measurement error in surveys: A review. *Journal of Official Statistics, 16*(4), 331-361.

Morse, J. M., & Niehaus, L. (2009). *Mixed method design: Principles and procedures*. Walnut Creek: Left Coast Press.

Muthén, B. (2006). The potential of growth mixture modeling. *Infant and Child Development, 15*

Muthén, B. (2010). *Bayesian analysis in mplus: A brief introduction* No. version 3). Los Angeles: Muthen & Muthen.

Muthén, L. K., & Muthén, B. (2010). *MPLUS*. Los Angeles, CA:

Muthén, L. K., & Muthén, B. (2010). *MPLUS user's guide*. Los Angeles, CA: Muthén & Muthén.

Nicoletti, C., & Peracchi, F. (2002). A cross-country comparison of survey nonparticipation in the ECHP. *ISER Working Paper,* 2002-32

Nylund, K. L., Asparouhov, T., & Muthén, B. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A monte carlo simulation study. *Structural Equation Modeling, 14*(4), 535-569.

Palmquist, B., & Green, D. P. (1992). Etsimation of models with correlated measurement errors from panel data. *Sociological Methodology, 22*, 119-146.

Pennell, S. G. (1993). Cross-sectional imputation and longitudinal editing procedures in the survey of income and program participation. *SIPP Working Paper 1993-06.*

R Core Development Team (2009). *R: A language and environment for statistical computing*, version 2.9.1.Vienna, Austria: www.r-project.org

Reeve, C. L., & Lam, H. (2005). The psychometric paradox of practice effects due to retesting: Measurement invariance and stable ability estimates in the face of observed score change. *Inltelligence, 33*(5), 535-549.

Rips, L. J., Conrad, F. C., & Fricker, S. S. (2003). Straightening the seam effect in panel surveys. *Public Opinion Quarterly, 67*(4), 522-554.

Rogelberg, S. G., Fisher, G. G., Maynard, D. C., Hakel, M. D., & Horvath, M. (2001). Attitudes toward surveys: Development of a measure and its relationship to respondent behavior. *Organizational Research Methods, 4*(1), 3-25.

Ryan, R. M., & Connell, J. P. (1989). Perceived locus of causality and internalization: Examining reasons for acting in two domains. *Journal of Personality and Social Psychology, 57*(5), 749-761.

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology, 25*, 54-67.

Sala, E., Uhrig, S. N., & Lynn, P. (2011). "It is time computers do clever things!" the impact of dependent interviewing on interviewer burden. *Field Methods, 23*, 3-23.

Saris, W. E., & Andrews, F. (1991). Evaluation of measurement instruments using a structural modeling approach. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz & S. Sudman (Eds.), *Measurement errors in surveys* (pp. 575-597). New York: Wiley.

Saris, W. E., & Gallhofer, I. N. (2007). *Design, evaluation and analysis of questionnaires for survey research*. New York: Wiley.

Scherpenzeel, A., & Das, M. (2011). "True longitudinal and probability-based internet panels: Evidence from the Netherlands. In M. Das, P. Ester & L. Kaczmirek (Eds.), *Social and behavorial research and the interne - advances in applied methods and research strategies* (pp. 77-104). New York: Routledge.

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review, 18*, 210-222.

Schonlau, M., van Soest, A., Kapteyn, A., & Couper, M. (2009). Selection bias in web-surveys and the use of propensity scores. *Sociological Methods & Research, 37*(3), 291-318.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis - modeling change and event occurrence*. New York: Oxford University Press.

Smyth, J. D. (2006). Comparing check-all and forced-choice question formats in web surveys. *Public Opinion Quarterly, 70*(1), 66.

Statistics-Netherlands. (2009). *Statline database*. Voorburg: Statistics Netherlands, last visisted on the 28th of July 2009, http://statline.cbs.nl

Steenkamp, J. ., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research, 25*(1), 78-107.

Stocké, V. (2006). Attitudes toward surveys, attitude accessibility and the effect on respondents' susceptibility to nonresponse. *Quality and Quantity, 40*, 259-288.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (fifth ed.). New York: Pearson Education Inc.

Taylor, M. F., Brice, J., Buck, N., & Prentice-Lane, E. (2009). *British household panel survey user manual volume A: Introduction, technical report and appendices.* Colchester: University of Essex.

Terborg, J. R., Howard, G. S., & Maxwell, S. E. (1980). Evaluating planned organizational change: A method for assessing apha, beta, and gamma change. *Academy of Management Review, 5*(1), 109-121.

Toepoel, V. (2008). Effects of design in web surveys. *Public Opinion Quarterly, 72*(5), 985.

Tortora, R. D. (2009). Attrition in consumer panels. In P. Lynn (Ed.), *Methodology of longitudinal surveys* (pp. 235-248). Chichester: Wiley.

Tourangeau, R., Groves, R. M., & Redline, C. D. (2010). Sensitive topics and reluctant respondents. demonstrating a link between nonresponse bias and measurement error. *Public Opinion Quarterly, 74*(3), 413-432.

Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2000). *The psychology of survey response*. Cambridge ; New York: Cambridge University Press.

Tuten, T. L., & Bosnjak, M. (2001). Understanding differences in web usage: The role of need for cognition and the five factor model of personality. *Social Behavior and Personality, 29*(4), 391-398.

Uhrig, S. N. (2008). The nature and causes of attrition in the British household panel survey. *ISER Working Paper,* (5)

Vallerand, R. J., & Blisonnette, R. (1992). Intrinsic, extrinsic and amotivational styles as predictors of behavior: A prospective study. *Journal of Personality, 60*(3), 599-620.

Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods, 5*(2), 139-158. doi:10.1177/1094428102005002001

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance Literature: suggestions, practices and

recommendations for organizational research. *Organizational Research Methods, 3*(4), 1-70.

Vandenberg, R. J., & Self, R. M. (1993). Assessing newcomers' changing commitments during the first 6 months of work. *Journal of Applied Psychology, 78*(4), 557-568.

Voogt, R. J. J., & Saris, W. E. (2005). Mixed mode designs: Finding the balance between nonresponse bias and mode effects. *Journal of Official Statistics, 21*, 367-387.

Voorpostel, M. (2009). Attrition in the Swiss household panel by demographic characteristics and levels of social involvement. *FORS Working Paper, 1_09*

Wagner, G. G., Frick, J.,R., & Schupp, J. (2006). Enhancing the power of household panel studies - the case of the German socio-economic panel study. Paper presented at the *StatCan Conference "Longitudinal Social Surveys in an International Perspective",* Montreal.

Watson, N., & Wooden, M. (2009). Identifying factors affecting longitudinal survey response. In P. Lynn (Ed.), *Methodology of longitudinal surveys* (pp. 157-182). Chichester: Wiley.

Webber, M. (1994). The survey of labour and income dynamics: Lessons learned in testing. *SLID Research Paper,* (07)

Wiley, J. A., & Wiley, M. G. (1974). A note on correlated errors in repeated measurements. *Sociological Methods & Research, 3*, 172-188.

Wiley, D. E., & Wiley, J. A. (1970). The estimation of measurement error in panel data. *American Sociological Review, 35*, 112-117.

Zaller, J. (1992). *The nature and origins of mass opinion*. Cambridge: Cambridge University Press.

Appendix A: additional tables for the propensity matching procedures

Table A 1: T-tests for differences between matched and unmatched samples

T-values (df)	Before matching WAPI-panel	Matched samples WAPI-panel	Unmatched samples WAPI-panel
Dust Industry	**-9.19 (917)**	**-1.97 (368)**	**-8.73 (533)**
Bad smell industry	**-11.41 (968)**	**-3.65 (369)**	**-10.30 (558)**
Noise Industry	**-6.75 (928)**	-1.23 (371)	**-6.28 (528)**
Bad smell traffic	**-2.11 (1798)**	0.90 (370)	**-2.10 (1404)**
Noise traffic	**-2.67 (1806)**	**-2.00 (371)**	**-3.16 (1411)**
Noise airplanes	-0.82 (1807)	0.44 (372)	-1.35 (1411)
Light pollution	-0.44 (1801)	-0.21 (372)	-0.10 (1405)
Mean Score 7 items	**-6.23 (1818**	-0.62 (372)	**-5.91 (1422)**
Significant difference	**6/8**	3/8	6/8

- df are rounded to nearest number
Statistics in **bold**: significant with p<0.05

Table A 2: T-tests for differences between matched and unmatched samples

T-values (df)	Before matching CATI-WAPI	Matched samples CATI-WAPI	Unmatched samples CATI-WAPI
Dust Industry	**11.37 (2435)**	**5.25 (1744)**	**10.61(976)**
Bad smell industry	**12.58 (2405)**	**5.49 (1443)**	**11.80 (965)**
Noise Industry	**12.15(2225)**	**6.69 (1331)**	**10.17 (1005)**
Bad smell traffic	**8.86 (2523)**	**3.90 (1486)**	**8.53 (897)**
Noise traffic	**9.87 (2590)**	**4.95 (1485)**	**8.95 (1049)**
Noise airplanes	**9.01 (2271)**	**3.81 (1414)**	**8.28 (889)**
Light pollution	**9.73 (2233)**	**3.98 (1397)**	**8.06 (873)**
Mean Score 7 items	**14.85 (2358)**	**7.06 (1422)**	**13.48 (1942)**
Significant difference	8/8	8/8	8/8

-df are rounded to nearest number
Statistics in **bold**: significant with p<0.05

Appendix B: Effects of DI and edit checks on reporting in experimental and BHPS data

In the experimental data 7.59% of all cash transfer sources were reported in response to the RDI follow-up questions (Table B 1). In the BHPS data at waves 16 and 17, about 2% of all cash transfer sources were reported in response to the edit checks, and about 7 % in response to the RDI follow-up question. The total proportion of cash transfer sources reported in response to edit checks and DI are therefore similar in the BHPS and the experimental data. For the other income types for which DI was used at waves 16 and 17 in the BHPS, the proportion of income sources reported in response to the RDI follow-up was around 5% for private pensions, around 9% for other transfers and around 11% for investment income. DI therefore had similar effects for these other income types as for cash transfer income.

Table B 1: Number and Percentage of Income Sources Reported in Response to INDI, Edit Check and RDI, by Income Type, Wave, and Data Source

		Cash transfers		Pensions		Other transfers		Investment	
		N	%	N	%	N	%	N	%
BHPS-Wave15	INDI	8088	98.58	1717	100.00	426	100.00	274	100.00
	Checks	117	1.42	–	–	–	–	–	–
Wave16	INDI	8170	91.28	1776	93.62	515	90.35	323	89.23
	Checks	165	1.84	–	–	–	–	–	–
	RDI	615	6.87	121	6.38	55	9.65	39	10.77
Wave17	INDI	7895	92.25	1846	95.15	501	91.09	302	87.54
	Checks	157	1.83	–	–	-	–	–	–
	RDI	506	5.91	94	4.85	49	8.91	43	12.46
Experiment	INDI	487	92.41	–	–	–	–	–	–
	RDI	40	7.59	–	–	–	–	–	–

Notes: N are counts of the total number of income sources reported in a wave, over all original BHPS respondents. N respondents wave 15:8,538,.wave 16:8,484, wave 17:8,322, experimental data: 274.
% are column percentages within a wave.

At the household level, 100% of cash transfer sources reported in response to the RDI follow-up in the experimental data were sources that had not already been reported by another household member (Table B 2). In the BHPS data at waves 16 and 17, 77.6% and 74.1% of cash transfer sources reported in response to the RDI, and 86.7% and 88.6% of sources in response to the edit checks were unique reports. This suggests that while the percentage of unique reports is somewhat lower in the BHPS than the experimental data, the general trend is similar. The effect of the edit checks was again in the same direction as the effect of RDI. For private pensions the percentage of sources reported in response to RDI,

that had not already been reported by another household member, was 94.1% and 100%; for other transfers 100% and for investment income 63.6% and 64.3%. DI therefore had similar effects for these additional income types as for cash transfers. Both DI and edit checks increased the number of unique cash transfers reported at the household level, for all types of non-labor income.

Table B 2: Mean Number of Total and Unique Income Sources Reported by Households, in Response to INDI, Edit Check and RDI, by Income Type, Wave, and Data Source

		Cash transfers			Pensions			Other transfers			Investment		
		Mean	Unique	% u	Mean	Unique	% u	Mean	Unique	% u	Mean	Unique	% u
BHPS-	INDI	2.266	2.001	88.3	.481	.477	99.2	.119	.114	95.8	.077	.051	66.2
Wave15	Checks	.033	.023	70.0	–	–	–	–	–	–	–	–	–
Wave16	INDI	2.422	2.117	87.4	.522	.517	99.0	.157	.150	95.5	.100	..063	63.0
	Checks	.045	.039	86.7	–	–	–	–	–	–	–	–	–
	RDI	.170	.132	77.6	.034	.032	94.1	..015	.015	100	.011	.007	63.6
Wave17	INDI	2.405	2.064	85.8	.546	.540	98.9	.154	.148	96.1	.097	.058	59.8
	Checks	.044	..039	88.6	–	–	–	–	–	–	–	–	–
	RDI	.143	.106	74.1	.005	.005	100.0	.015	.015	100	.014	.009	64.3
Experiment	INDI	2.706	2.578	95.2	–	–	–	–	–	–	–	–	–
	RDI	.222	.222	100.0	–	–	–	–	–	–	–	–	–

Notes: Mean refers to the mean number of sources reported per household, including households which did not report any income of a particular source. Unique excludes duplicate reports of the same income source by multiple household members. % u indicates the percentage of total reports per household that are unique reports. Number of respondent households at wave 15: 3570, wave 16: 3628, wave 17: 3558, experimental data: 180.

In the experimental data the amounts for 12.5% of cash transfer incomes reported in response to RDI had already been reported as part of a different source (Table B 3). This is a bit higher than the percentage in the BHPS cash transfer data at wave 17 (9.5%), but lower than at wave 16 (18.5%). The effect of RDI therefore seems to be similar for the reporting of cash transfer amounts in the BHPS and the experimental data. For the edit checks in the BHPS, the percentage of cash transfer sources for which amounts had already been reported is lower at 5.5% and 4.5%. For the other income types the percentage of RDI reports with zero amounts are also lower, at 9.1% and 5.3% for private pensions, 5.5% and 0% for other transfers, and 5.1% and 2.3% for investment income. The effects of edit checks for cash transfer income and of RDI for the other income types in the BHPS data therefore appear similar to the effects of RDI in the experimental data, in that the majority of additional income reports are associated with additional amounts.

Table B 3: Percentage of Reported Sources with Income Amounts of Zero, in Response to INDI, Edit Check and RDI, by Income Type, Wave, and Data Source

		Cash transfers	Pensions	Other transfers	Investment
BHPS-Wave15	INDI	6.0	1.0	0.5	1.5
	Checks	8.5	–	–	–
Wave16	INDI	8.2	1.3	0.7	1.7
	Checks	5.5	–	–	–
	RDI	18.5	9.1	5.5	5.1
Wave17	INDI	6.3	0.9	0.5	1.2
	Checks	4.5	–	–	–
	RDI	9.5	5.3	0.0	2.3
Experiment	INDI	4.7	–	–	–
	RDI	12.5	–	–	–

Notes: % is the percentage of respondents who reported receipt of an income source, but said the amount had already been included elsewhere. Based on all income sources reported by respondents as documented in Table B .

Table B 4: Mean Months of Receipt in Response to INDI, Edit Check and RDI, by Income Type, Wave, and Data Source

		Cash transfers	Pensions	Other transfers	Investment
BHPS-Wave 15	INDI	10.7	11.3	7.0	9.7
	Checks	10.6	–	–	–
Wave 16	INDI	10.8	11.0	7.8	9.7
	Checks	8.6	–	–	–
	RDI	9.9	10.5	8.3	9.1
Wave 17	INDI	10.7	11.0	7.5	10.2
	Checks	7.5	–	–	–
	RDI	10.0	9.7	8.8	9.0
Experiment	INDI	11.5	–	–	–
	RDI	9.6	–	–	–

Notes: The maximum duration of a spell is 12 months (1 Sept. – 1 Sept.). If a respondent reported two spells of receipt of the same income type within one year, the total number of months of receipt is counted, regardless of whether these were part of only one or multiple spells. Based on the number of income sources as documented in Table B

In both the experimental data and the BHPS, the mean duration of receipt of cash transfers is around 11 months in the INDI data, and around 10 months in the RDI data (Table B 4). The effect of RDI on reporting of duration of receipt therefore appears similar in the BHPS and experimental data. With edit checks the mean duration of receipt for BHPS cash transfer income is slightly lower, at around 9 months. For the other income types, the mean durations in the INDI data are around 11 months for pensions, 7 to 8 months for other transfers and around 10 months for investment income. In each case, the mean duration of

receipt with DI is similar to that with INDI. This suggests that the effect of DI on reporting of durations of cash transfer receipt is similar in the experimental and the BHPS data. The effects of the edit checks go in the same direction, and the effects of DI for the other income types are also similar to those for cash transfer income.

In both the experimental data and the BHPS data, RDI increases the percentage of transitions onto receipt and of continued receipt, while it decreases the percentage of transitions off cash transfer receipt and of continued non-receipt (Table B 5). The effects of edit checks in the BHPS cash transfer data are similar for all transition types, as are the effects of RDI for the other income types. Although all effects are small, we conclude that the effects of DI are similar in the BHPS and experimental data, that the effects of the edit checks go in the same direction, and that the effects of DI on the other types of non-labor income are similar to the effects for cash transfer income.

Table B 5: Transitions onto and of Income Receipt between Waves in Response to INDI, Edit Check and RDI, by Income Type, Wave, and Data Source

			Wave *t* receipt status					
Wave *t-1* receipt status			INDI		INDI + checks		INDI + checks + RDI	
Income	Waves		0	1	0	1	0	1
Cash transfers	15-16	0	94.53	1.07	94.19	1.41	–	–
		1	0.90	3.50	0.81	3.58	–	–
	16-17	0	94.47	1.00	94.13	1.33	93.99	1.47
		1	0.96	3.57	0.90	3.64	0.25	4.29
	Experiment*	0	90.53	2.11	–	–	90.49	2.14
		1	1.39	5.97	–	–	0.76	6.60
Pensions	16-17	0	92.02	0.95	–	–	91.86	1.08
		1	1.70	5.33	–	–	1.32	5.74
Other transfers	16-17	0	98.84	0.38	–	–	98.80	0.42
		1	0.40	0.38	–	–	0.18	0.60
Investment	16-17	0	97.58	0.62	–	–	97.52	0.68
		1	0.62	1.18	–	–	0.16	1.64

Notes: Numbers are cell percentages for each 2 by 2 transition matrix. Receipt status 0: no receipt, 1:receipt. * No edit checks were used in the experiment, therefore the final two columns refer to INDI + RDI in the case of the experimental data. Based on the number of respondents multiplied by the number of potential income sources. Number of respondents for transitions between waves 15 and 16: 7,820 , waves 16 and 17: 7,709, experimental data: 274.

Appendix C. Questions used for measuring study motivations

Study motivation

People have different reasons to do their best for their education. Looking at yourself, what are your reasons for studying? I study because …

Response scale 1. Not true at all, 2. Not true, 3. Sort of true, 4. True, 5. Very true

External motivations

1. I'll get into trouble if I don't
2. That's what I'm supposed to do
3. I don't want my teachers to get angry with me
4. It's the right thing
5. I don't want people to get mad at me
6. I want my teachers to think I'm a good student
7. I feel bad if I don't
8. I feel ashamed if I don't
9. I want other students to think I'm smart
10. It bothers me if I don't
11. I want people to like me

Internal Motivations

12. I want to understand the subjects
13. I want to learn new things
14. I want to find out if I'm right or wrong
15. I think it's important
16. I want it
17. It's fun
18. I enjoy it

Code system for analysis of qualitative interviews

Transition (5)

You have to do it yourself (3)

- Uncertainty (6)
- Professional future (3)

Needs (0)

- Advanced insight (1)
- Growth in learning (6)
- Critical thinking (4)
- Social contacts (2)

Reasons studying psychology (9)

- Going to university (1)
- All I ever wanted (2)
- Broad interest (1)

Decisiveness (1)

- Realised expectations (1)
- Decision specialization track (6)
- Purposefulness activities (6)
- Reassurance (1)
  - Pleasure (3)
  - Interesting (6)
  - I can do it (2)

Appendix D: growth parameters for GMM-model

Table D 1: Standardized growth parameters of the Growth Mixture Model with 4 classes and a free slope and quadratic slope variance

Parameters/ Class	1 Fast attriters	2 Lurkers	3 Gradual attriters	4 Loyal stayers
I	-	-	-	-
S	2.85 (2.43)	.19 (.03)**	1.88 (.42)**	.67 (.31)*
Q	1.41 (0.45)**	-.39 (.03)**	-1.10 (.17)**	-1.79 (.22)**
S with Q	.00 (.00)	-.87 (.01)**	.62 (.04)**	-.04 (.01)**

Notes: the intercepts are not estimated
p <0.05, ** p<0.01

List of Figures

List of Tables

Samenvatting (summary in Dutch)

Dit proefschrift gaat over de methodologie van panel surveys. In panel surveys worden dezelfde mensen door de tijd heen gevolgd. Het doel van dit soort onderzoek is het meten van ontwikkeling of verandering: bijvoorbeeld bij scholieren, patiënten, of huishoudens.

Het meten van veranderingen door de tijd heen kan alleen op een goede manier als er geen fouten in het onderzoek ontstaan. Het bestuderen van de ernst van fouten in onderzoek is het onderzoeksterrein van methodologen. In een aantal fases van het onderzoek kunnen fouten worden geïntroduceerd: bij het trekken van een steekproef voor de onderzocht populatie (dekkingsfout), wanneer potentiële respondenten niet willen meedoen (nonresponsfout) en in het stellen van vragen aan respondenten (meetfouten). In panelonderzoeken komen deze fouten op dezelfde manier voor als in eenmalige onderzoeken, maar de fouten hebben ook een specifieke longitudinale component. Zo kunnen respondenten er in elke fase van het panelonderzoek voor kiezen om niet langer mee te doen: paneluitval (attritie). Ook kunnen meetfouten er voor zorgen dat twee opeenvolgende metingen niet bruikbaar zijn om veranderingen goed te meten. Wanneer we een verschil meten tussen twee metingen, wordt die dan veroorzaak door echte verandering, of door meetfouten? Vijf studies in dit proefschrift beschrijven verschillende methoden om inzicht te krijgen in de grootte van surveyfouten in panelonderzoek. Iedere studie behandelt een specifieke soort fout, en in enkele hoofdstukken wordt ook onderzocht hoe verschillende surveyfouten met elkaar samenhangen. Fouten kunnen met elkaar samenhangen door een gemeenschappelijke oorzaak voor beide fouten. Een voorbeeld hiervan is de samenhang tussen nonrespons- en meetfouten die kan worden veroorzaakt door de motivatie van respondenten om mee te doen aan het doen. Wanneer respondenten niet gemotiveerd zijn om mee te doen, kan dat leiden tot nonresponsfouten, omdat respondenten weigeren mee te doen. Wanneer een respondent echter niet weigert, kan een lage motivatie echter ook leiden tot meetfouten in het onderzoek, omdat een respondent zo min mogelijk moeite doet om vragen te beantwoorden. Het begrijpen van de samenhang tussen bijvoorbeeld nonrespons- en meetfouten is belangrijk: op dit moment proberen survey-onderzoekers mensen te overtuigen om mee te doen aan surveys door het sturen van herinneringen en het geven van beloningen. We weten echter niet of mensen die moeten worden overgehaald betere of slechtere antwoorden geven dan mensen die niet overtuigd hoeven te worden. Het onderzoeken van de samenhang en afwegingen tussen verschillende

typen van surveyfouten is lastig, omdat ze altijd gelijktijdig voorkomen en lastig te isoleren zijn. In dit proefschrift worden statistische modellen voorgesteld die het onderzoeken en corrigeren voor surveyfouten mogelijk maken.

In het derde hoofdstuk wordt ingegaan op twee typen surveyfouten die ontstaan door de mode van dataverzameling. Surveys kunnen worden gedaan met behulp van een interviewer of zonder interviewer, en op papier, Internet of telefoon. Elke dataverzamelingsmethode heeft voor- en nadelen. Zo kan een papieren vragenlijst gemakkelijk per post worden verzonden naar iedereen die binnen de steekproef valt, terwijl in telefonische onderzoeken mensen zonder vaste telefoonlijn niet bereikt kunnen worden. Aan de andere kant is de nonrespons meestal hoger in postsurveys dan in telefonische surveys. De mode van dataverzameling is dus de gemeenschappelijke oorzaak van verschillen in dekkingsfouten, nonresponsfouten en meetfouten.

Het is tegenwoordig gebruikelijk om meerdere dataverzamelingsmethodes tegelijk of na elkaar in te zetten in surveys, om te zorgen voor een zo hoog mogelijke respons en om de kosten in de hand te houden. Hoe de resultaten uit de verschillende dataverzamelingsmethodes gecombineerd moeten worden, is onduidelijk. Verschillen in de resultaten tussen de dataverzamelingsmethodes kunnen zijn veroorzaakt door verschillen in het type mensen dat meedoet (dekkings- en nonresponsfouten) of de manier waarop mensen verschillend reageren op enquêtevragen (meetfouten). Deze verschillen tussen de resultaten uit verschillende dataverzamelingsmethodes wordt het mode-effect genoemd.

In het derde hoofdstuk wordt propensity score matching gebruikt om verschillende typen surveyfouten te isoleren. Met behulp van matching worden twee respondenten uit een mixed-mode onderzoek dat wordt afgenomen via Internet of telefoon aan elkaar gekoppeld. Deze koppeling gebeurt wanneer de respondenten sterk op elkaar lijken op een set met achtergrondvariabelen. Dat wil zeggen dat na matching, er sets van gekoppelde respondenten uit iedere methode bestaan, die sterk op elkaar lijken. Verschillen in dekkings- en nonresponsfouten zijn na matching voor deze variabelen dus verdwenen. Uit het onderzoek blijkt dat na matching de verschillen in resultaten tussen de Internet en telefonische methodes kleiner worden, maar niet verdwijnen. Het feit dat de verschillen niet verdwijnen zou kunnen komen doordat de achtergrondvariabelen die zijn gebruikt niet voldoende de verschillen tussen de samenstelling van de twee steekproeven verklaren. Waarschijnlijker is dat de verschillen tussen de Internet en telefonische survey na matching duiden op een verschil in meetfouten tussen de methodes. In een internetsurvey wordt met behulp van dezelfde

enquêtevragen iets anders gemeten dan in een telefonische enquête. Deze bevinding impliceert dat het gevaarlijk is om resultaten uit mixed-mode surveys samen te voegen. Het lijkt vooralsnog beter te zijn om de resultaten uit verschillende dataverzamelingsmethodes apart te rapporteren. In ieder geval totdat we meer inzicht hebben in de exacte verschillen in meetfouten tussen enquêtemethodes.

In het vierde en vijfde hoofdstuk van dit proefschrift wordt ingegaan op meetfouten die ontstaan bij het meten van veranderingen door de tijd heen. Door meetfouten in iedere meting, kunnen veranderingen of ontwikkelingen door de tijd heen niet goed geschat worden. Wanneer iemand in achtereenvolgende metingen een jaarlijks netto-inkomen rapporteert van €34.000 en €36.000, is er dan sprake van inkomensgroei, of van meetfouten in de metingen?

Een methode om longitudinale meetfouten te verminderen is het gebruik van Dependent Interviewing (DI). Met DI worden data uit een eerdere meting gebruikt in het interview van de latere meting. De respondent krijgt actief zijn antwoorden uit het vorige interview te zien of horen. Hierna wordt er in verschillende varianten van Dependent Interviewing gevraagd of “dat nog zo is” er “verandering heeft plaatsgevonden” of “wat voor antwoord de respondenten nu wil geven”. Ook kan de data uit eerdere interviews niet actief gebruikt worden, maar alleen wanneer data uit twee metingen niet overeen lijken te komen.

Dependent Interviewing is eerder gebruikt in panelonderzoeken, maar het is onduidelijk wat het effect van DI nu is op datakwaliteit. Er zouden namelijk ook nadelen kunnen zitten aan het gebruik van DI. Respondenten kunnen gemakkelijk hun oude antwoord bevestigen, zonder na te denken, of hun oude antwoord als waarheid beschouwen, zelfs als dat antwoord eigenlijk niet helemaal klopte. In hoofdstuk 4 wordt door middel van een experiment bekeken welke vorm van Dependent Interviewing de beste datakwaliteit oplevert. Om de datakwaliteit te beoordelen worden inkomensdata van vier opeenvolgende metingen gebruikt in een statistisch model dat de betrouwbaarheid van de metingen in elke experimentele conditie schat. In hoofdstuk vijf wordt gekeken wat de gevolgen zijn van DI voor het meten van veranderingen in inkomen voor huishoudens in het Verenigd Koninkrijk. Uit beide studies blijkt dat de invoering van DI een positief effect heeft op de kwaliteit van inkomensdata. Het actief terugkoppelen van antwoorden uit een vorig interview blijkt minder goed te werken dan het gebruiken van de data als controle-instrument, wanneer er sprake lijkt te zijn van grote veranderingen of inconsistenties.

In hoofdstuk zes wordt dieper ingegaan op het meten van veranderingen met behulp van antwoordschalen. Zulke antwoordschalen worden veel gebruikt door psychologen en sociologen om complexe

begrippen te meten. In hoofdstuk 6 wordt als voorbeeld het meten van studiemotivatie onder eerstejaars studenten gebruikt. Bij het gebruik van schalen, zoals voor studiemotivatie wordt aangenomen dat het begrip 'studiemotivatie' hetzelfde betekent voor studenten gedurende de studie. Deze stabiliteit kan worden gemeten door te kijken naar de grootte van factorladingen op de verschillende meetmomenten. Vaak, en zo ook voor studiemotivatie, blijkt echter dat de factorladingen over de tijd heen verschillen. Studiemotivatie betekent voor studenten wezenlijk iets anders, wanneer ze aan het begin van hun studie, en aan het eind van hun eerste jaar dezelfde vragenlijst voorgelegd krijgen. Uit enquêtestudies is het lastig om te begrijpen waarom dit gebeurt en wat leidt tot een veranderende kijk op studiemotivatie. Daarom zijn tegelijk met de enquêtes enkele studenten geïnterviewd. Tijdens die interviews werd dieper ingegaan op hun ervaringen om te studeren, en de moeilijkheden en veranderingen die ze daarbij ervoeren.

De informatie die uit de kwalitatieve interviews kwam convergeerde met de informatie uit de enquête en verklaarde waarom bepaalde enquêtevragen door studenten anders geïnterpreteerd werden. Studenten konen aan het begin van hun studie bijvoorbeeld niet goed evalueren of ze studeerden, omdat ze dat interessant vonden. Aan het eind van het eerste jaar bleek deze vraag veel belangrijker voor het meten van het begrip studiemotivatie. Door het toevoegen van open interviews aan een panelonderzoek kunnen veranderingen op micro- en macro niveau beter begrepen worden.

Het laatste hoofdstuk in dit proefschrift behandelt het probleem van paneluitval. Eerdere studies naar uitval vergeleken meestal de uitvallers uit een studie met de blijvers om te zien of de uitvallers verschilden van de blijvers. In veel panelsurveys is het uitvalproces echter veel subtieler, zeker wanneer er frequent data worden verzameld. Sommige respondenten zullen altijd mee doen, anderen na een paar keer nooit meer. Er zijn echter ook respondenten die vaak meedoen, maar soms een meetmoment overslaan, of respondenten die af en toe meedoen. Het is van belang om dit onderscheid te maken, omdat het waarschijnlijk is dat elke groep uitvallende respondenten van elkaar verschilt. Respondenten die vinden dat hun privacy wordt geschonden zullen resoluut stoppen, terwijl respondenten die niet erg gemotiveerd zijn onregelmatig zullen meedoen. Het bestuderen van het attritieproces, en de voorspellers daarvan kan informatie opleveren die interessant kan zijn voor panelmanagers.

In het hoofdstuk wordt een Latente Klasse Strutureel Vergelijkings Model gebruikt om respondenten te classificeren in verschillende groepen uitvallers. Trouwe respondenten (die altijd meedoen aan onderzoeken) blijken ouder en consciëntieuzer te zijn, en respondent die

onregelmatig meedoen blijken jonger en lager opgeleid te zijn. Het hoofdstuk eindigt met het bespreken hoe iedere groep uitvallers bijdraagt aan meetfouten in stemgedrag. Elke groep blijkt een specifiek stemgedrag te hebben dat afwijkt van het stemgedrag van het hele panel.

About the author

Peter Lugtig (1983) studied political science at the University of Amsterdam and Uppsala University, specializing in European politics and methods and statistics. Since 2006 he works at the department of methods and statistics of Utrecht University, where he specializes in survey methodology. His areas of expertise include questionnaire design, nonresponse analysis and analysis of longitudinal data. Apart from teaching on these subjects, Peter worked on a dissertation project under supervision of prof. dr. Joop Hox and prof. dr. Gerty Lensvelt-Mulders. The five studies that comprise this dissertation are the product of that project. Several chapters in this book have been published, or will soon be published as separate research articles.

In spring 2009, Peter Lugtig conducted a part of his research at the Institute of Economic and Social Research, where he worked with Annette Jäckle on a paper discussing the use of Dependent Interviewing and edit checks in the British Household Panel Survey. From September 2011, Peter will continue teaching and doing research at the department of methods and statistics at Utrecht University.

Publications

Lugtig, P., Boeije, H.R. and Lensvelt-Mulders, G.J.L.M. (in press) Change? What change? using mixed-methods research to understand longitudinal measurement invariance*, Methodology*

Lugtig, P., Lensvelt-Mulders, G.J.L.M., R. Frerichs and Greven, A. (2011), Estimating nonresponse bias and mode effects in a mixed-mode survey*, International Journal of Market Research,* 53(5).

Lugtig, P. and Jäckle, A (in press) In-Interview edit checks: effects on measurement error in non-labour income and estimates of household income and poverty, I*SER working paper series 2011-23.* Colchester: Institute for Social and Economic Research, University of Essex

De Ridder, D.T.D, de Boer, B.J., Lugtig, P. and Bakker, A. (2011*)* Not doing bad things is not equivalent to doing the right thing: Distinguishing between inhibitory and initiatory self-control, *Personality and Individual Differences,* advance access

Lensvelt-Mulders, G.J.L.M., Lugtig,. P. and Hubregtse, M. (2009) Separating Selection bias and Non-coverage in Internet Panels using Propensity Matching, *www.surveypractice.org,* August 2009

Lensvelt-Mulders, G.J.L.M., Hox, J.J. and Lugtig, P. (2008) Assembling an access panel: a study of initial nonresponse and self-selection bias, in Stoop, I and Wittenberg, M. (eds.) *Access panels and online research, panacea or pitfall?* Amsterdam, Aksant Publishers.